

A Hybrid Methodology Combining Statistical Learning and Econometrics: Forecasting Algeria's Real GDP Using LASSO-OLS Algorithmic Approach

Mohamed BOUZIANE¹, Belkacem BOUDALI², Redouane GUELAILIA³

¹ University of Tissemsilt, Laboratory of Modern Economic and Sustainable Development, Algeria,
mohamed.bouziane@univ-tissemsilt.dz

² University of Tissemsilt, Algeria, prof.boudali.belkacem@gmail.com

³ University of Tissemsilt, Algeria, guelailia.redouane@univ-tissemsilt.dz

Received: 06/07/2025

Accepted: 11/08/2025

Published: 15/09/2025

Abstract:

This study explores the integration of statistical learning and OLS, intending to model the Real GDP over Algeria macroeconomic data.

For this ending, we use sequential algorithm. At first step, algorithm extends the macroeconomic dataset by performing data transformation, and then it pools the dataset into multiple folds, each fold contains all dataset excluding one variable. Variable selection and OLS regression are iterated for each fold, subjecting resulting models to accuracy tests. Model with insignificant feature(s) is banned from model selection, else the model exhibiting the lowest cross-validated RMSE is chosen for RGDP prediction.

Despite model interpretability, our proposed methodology demonstrates its effectiveness for conducting accurate results. Consequently, the model performed to predict Algeria's Real GDP achieves high-accuracy, consistently establishing 99.99% of goodness of fit and meeting several linear regression assumptions.

The findings provide valuable insights into economic performance mechanisms, particularly within the Algerian context. These insights may also be generalizable to other cross-sectional datasets and applicable for modeling various macroeconomic aggregates.

Keywords: Statistical Learning; Econometrics; Data Processing; LASSO; Variable Selection; Regression; Real GDP.

JelClassificationCodes: C51; C52; C53; E17; E31; E37.

1. INTRODUCTION

The main subject of econometrics is predicting macroeconomic indicators (Briggs, 2022), measuring relationships (Vuppapapati, 2021), and inferring the causality (Varian, 2014; Liang, 2021). The econometric approach employs economic theories to formulate hypotheses and derive empirical results (James, Witten, Hastie, & Tibshirani, 2013). The essence of econometrics is lying in validating these theories through data analysis, while ensuring that theoretical insights are empirically supported.

In contrast with practice; both theories and data tend generally to be heterogeneous (Stigum, 2003). Long time ago, in his writings “*on the definition and method of political economy*”, **John Stuart Mill** had noted the gap between theory and practice, as he wrote “*The most universal of the forms in which this difference of method is accustomed to present itself, is the ancient feud between what is called theory, and what is called practice or experience*” (Mill, 2007). Subsequently, numerous studies have experimented misspecification problems in econometric models, which is due from restrictive economic assumptions that reduces the accuracy of a model (Cerreia-Viogliao, Hansen, Maccheronia, & Marinacci, 2022). When explanatory variables exhibit spurious correlations with the response variable without causal relationships, conventional econometric models may be faded and necessitating a broader approach (Sala-i-Martin, 1997). In such situations, a rigorous economic analysis is needed to explore the divergences and opportunities across countries (Sakiru, Gil-Alana, & Gonzalez-Blanch, 2022). For instance, when analysing Gross Domestic Product (GDP), it is crucial to account for the complexity of income disparities (Sala-i-Martin, 1997), resources devoted to production (Barro & Sala-i-Martin, 2004), and the comprehensive understanding of economic conditions (Mankiw, Macroeconomics, 2015).

Consequently, recent research has increasingly shifted toward machine learning approaches based on data patterns rather than traditional theoretical-driven methods. This shift aligns with advancements in statistical learning, providing effective economic models. Controversy, conventional approaches fail to capture complex dynamics inherent in real-world data (Binner, Elger, Nilsson, & Tepper, 2004; Marwala, 2013). For instance, non-linear models offer more robust and reliable insights (James, Witten, Hastie, & Tibshirani, 2013). Whereas conventional models of econometrics have limited scale, requiring independent variables undergone to scale reduction (James, Witten, Hastie, & Tibshirani, 2013). In contrasts with models of statistical learning which can deal with large dataset without dimensionality reduction such in variable selection methods (Ratner, 2017).

Considering the issues and the advantages of econometrics and emerging skills in data science, this study seeks to investigate the benefits of applying statistical machine learning techniques and conventional regression methods for economic purposes. Accordingly, it focuses on leveraging data patterns to identify Algeria’s Real GDP trends. This approach could enable the derivation of real-time RealGDP insights, enhancing long-term economic sustainability. Therefore, the central research question of this study is: *What are the most relevant variables for modeling Algeria’s Real GDP by integration statistical learning methods?*

To address answering this question, the paper focus on reframing the path by exploring these three key features:

- What are the key driver features influencing economic performance, as reflected by the variables shaping Real GDP trends?
- How can statistical learning methods be integrated to identify these factors, given the structure and availability of the data?
- How effective are predictions based on the LASSO-OLS hybrid methodology in forecasting Real GDP?

First of all, to model Algeria's Real GDP, we have to introduce the most relevant variables that reflect its hydrocarbon-driven economy, demographic pressures, and macroeconomic dynamics. Therefore, key variables should include: hydrocarbon revenues (oil and gas exports), which dominate GDP due to their 19% direct contribution and 92% share of exports (International Monetary Fund, 2024). Government spending is also a major driver of demand via subsidies and public investment. Unemployment rate captures labour market inefficiencies (e.g., over 15% of unemployment in last 5 years). Inflation rate affects Real GDP through price pressures. And US Dollar Exchange Rate influences import costs and export earnings. Additional variables like population growth, foreign direct investment (FDI), commodity prices, and trade policy provide critical insights into accuracy improvements. These variables, including their non-linear patterns, are crucial for statistical learning, enabling to model Algeria's unique economic preferences.

Secondly, the employ of statistical learning presents a promising approach to perform models. Where variable selection techniques, which is introduced by James et al. (2013), serve powerful tools to enhance model accuracy.

Thirdly, by identifying the most influential variables for stimulating GDP growth in Algeria, a linear modeling approach proves appropriate in terms of efficiency, interpretability, and reduced complexity. Consequently, the LASSO-OLS methodology offers promising insights into predictive performance and model accuracy.

However, the significant aim of this paper is the development of an innovative approach to model Real GDP by integrating statistical learning methods with econometric techniques. Leveraging advanced computing techniques, the study aims also to improve the ability in inferring causality from large-scale datasets, resulting a more accurate economic model that effectively captures the real trends of Algerian GDP. This hybrid methodology links both conventional and new approaches, highlighting their combined value and figuring the relevance of such integrated framework for advancing economic analysis.

The motivation for this research stems from the growing prominence of artificial intelligence in recent studies and prompting the critics of traditional economic forecasting models. Unlike conventional methods, which rely on pre-defined predictor variables (Gujarati, 2011), advanced techniques need resource-intensive processes (Khan, et al., 2022; Ratner, 2017). Therefore, integrating statistical learning into economic analysis offers a promising alternative to both frameworks. This approach not only addresses limitations, which are discussed in the second section, but also aims to provide valuable insights into Algeria's distinct economic dynamics marked by: fluctuating oil revenues; extensive demographic trends; and a unique economic condition. By leveraging this innovative methodology, the study seeks to provide a

rapid, effective tool to foster sustainable growth and reviving Algeria's economy, bridging the gap between advanced computational techniques and practical economic challenges.

In the next chapter; we will introduce the theoretical framework of our methodology, followed by the third chapter, which details the proposed method. The fourth chapter discusses the results of this framework, and finally, the conclusion summarizes the key findings of the paper and its limitations for further insights.

2. THEORETICAL FOUNDATIONS AND RELATED WORK

Earlier economic modeling approaches are almost based on theoretical frameworks (Madurapperuma, 2023), which are generally geared to regression and rooted on conventional econometric foundations (Kohlscheen, 2021). On other hands modern techniques of econometrics are used in decision making, alongside correlation and regression in various literature. The use of empirical method in economics dated to the early seventeenth century (Charpentier, Flachaire, & Ly, 2018), later these techniques were called Econometrics. Econometric models usually assume that relationships are linear(Gujarati, 2011), whereas real world complexity violates this assumption (Marwala, 2013).

Machine learning is more recent, he was Arthur L. Samuel of IBM who established the first self-learning program (Patanjali, 2018), which is called later Machine Learning (ML). ML is a field of study that gives a computer the ability to make decision without being programmed (Kalali, Richerson, Ouzunova, & Westp, 2019). it is mostly founded on the theory of statistics, specifically geared towards modelling and able to learn. ML involves the use statistical tools to understand data, usually for the purposes of prediction or forecasting, amongst to solve an economic issue(Alpaydin, 2014). Otherwise, econometrics is the use of statistical tools to measure relationships(Gujarati, 2011) persuaded by economic theory(Ouliaris, 2011; Hayes, 2022). Later, many combinations arose to combine both fields, key among statistical Learning (SL).

2.1 Overview of Statistical Learning

Broadly speaking, statistical (machine) learning (or Inductive learning) is a subclass of machine learning that studies algorithms for learning knowledge based on statistical regularities (Sammur & Webb, 2017), a method to extract information and knowledge from data (Quintana, 2021), and a set of tools to understand data (James, Witten, Hastie, & Tibshirani, 2013).

There is two classification of SL methods; supervised or unsupervised (Quintana, 2021). The supervised statistical learning is the use of one or more inputs data to predict an output through a model. This is a subject of regression for purposes occur in economics and several disciplines. A sort in which there are inputs but outcome is not supervised or unpredictable, is called unsupervised learning; instead, we can learn relationships and structure from such data (James, Witten, Hastie, & Tibshirani, 2013).

The most valuable writing on statistical learning was articulated by Trevor Hastie; Robert Tibshirani and Jerome Friedman (2009). Their book "*The Elements of Statistical Learning*" collected the fruits of the methods and the model of statistical learnings. Later, an up-to-date version was published co-edited by Gareth James and Daniela Witten(2013) named "*An Introduction to Statistical Learning with Applications in R*" which is also formalized this field by

bending traditional statistics with machine learning techniques, making a foundation for data-driven analysis across disciplines, including economics. Matloff (2017) have made bridge between classical linear regression and modern machine learning models, providing R application. These books offer us as valuable resources for understanding statistical learning techniques and their implementation in R. The implementation of various statistical learning algorithms has shown promising results in enhancing the results of predictions. These algorithms have demonstrated significant improvements in optimizing processes, reducing costs, and increasing efficiency of prediction (Khan, et al., 2022).

2.2 Statistical Learning vs Econometrics

Applied econometrics uses data to evaluate economic theories (Gujarati & Porter, 2009) and to design models, thereby analysing economic history and forecasting future trends and outcomes. Likewise, SL is also able to analyse data and then find correlations, causations, and even predictions (Haponik, 2019). Nevertheless, SL is skilled to build more complex models on data (James, Witten, Hastie, & Tibshirani, 2013) which has been traditionally built around prediction (Chakraborty & Joseph, 2017), and through the accuracy of the model, the prediction assessments will be economically significant and statistically sounded (Reiss & Wolak, 2007). This technique is distinguished from econometrics by the high dimensional variables and data, alongside having opportunity to employ a large set of models in empirical research. Although, Econometrics and SL use statistical foundations and both approaches allow building predictive models (Flachaire, Hacheme, Hué, & Laurent, 2022), but Econometrics has been undergone rigorous respecting the economic theory (Ouliaris, 2011; Hayes, 2022), and has different methods (Charpentier, Flachaire, & Ly, 2018). Therefore, SL approaches have emerged as promising alternatives to these approaches, demonstrating the ability to analyze large scale datasets and identify complex patterns (Kuhn & Johnson, 2013).

Sometimes SL is considering as new brunch of Machine Learning (Sammut & Webb, 2017), because it depends on algorithms more alike for learning knowledge. On other occasions, this field was considered an improvement of regressions, which is in continues developing since long time ago (James, Witten, Hastie, & Tibshirani, 2013). Moreover, statistical learning and econometrics are both referring to a vast set of tools for understanding data (Hastie, Tibshirani, & Friedman, 2009), helping to build models by using data or data sets to make decisions, whereby being later implemented in computer programs. SL and Econometrics have a lot of limitations. Econometrics is basically statistical method geared toward answering economic questions, SL for economics has also a similar purpose but with the use of a massive amount of data and complex modeling approaches. Therefore, SL for economics is not based on exactly the same models as econometrics.

2.3 Related works

In the straight forwardness, our approach incorporates key principles and strategies derived from the insights provided in the context of both SL and econometric, ensuring a well-founded methodology to model Real GDP trends in context of Algeria macroeconomic.

Economic researchers rely on GDP as a tool to assess the extent of economic activity in a country, as noted by Mankiw (2015). By encompassing the total value of domestically produced goods and services, GDP provides a reliable measure of how efficiently an economy operates, reflecting its productive capacity and resource utilization. GDP presents an Omni-comprehensive of economic activities (Mankiw, 2015), because it deduces an important informative indicator of the economy wealth (Maccarrone, Morelli, & Spadaccini, 2021), and is relevant to predict the GDP for shaping policies and making decisions that drive macroeconomic stability and growth (Maccarrone, Morelli, & Spadaccini, 2021).

Meeting recent research and addressing those earning more marks for new approaches, several successful studies dealt with such topic. Likewise, studies have attempted to forecast Algeria's GDP using linear regression, several are based on different conventional timeseries approaches mainly: Ordinary Least Square (OLS) (Makhlouf, 2022), Vector Autoregressive (VAR) (Abdelmounaim & Abdelmounaim, 2024; Messaoudi, 2024; Messaoudi & Khouidmi, 2024), Autoregressive Distributed Lags (ARDL) (Smaili & Khelassi, 2021), alongside Machine Learning approaches, such as Random Forest (RF) (Messaoudi, 2024; Messaoudi & Khouidmi, 2024), Support Vector Regression (SVR) (Zermane, 2025), Gradient Boosting Machine (GBM) (Messaoudi & Khouidmi, 2024). Adaptive-Neuro-based Fuzzy Inference System (ANFIS) (Sahed, Kahoui, & Mekidiche, 2020). Other studies have employed various statistical learning models based on cross-sectional data from different regions, such as Support Vector Machine (SVM), Recurrent Neural Networks (RNN), and K-Nearest Neighbors (KNN) (Maccarrone, Morelli, & Spadaccini, 2021).

By examining literature of the SL uses in the economy, we find the interest is still continuing to grow rapidly and econometricians have devoted considerable researches on machine learning approach. For Instance, Hastie et al. (2009) introduced regularization technique (LASSO, Ridge and Elasticity Net regression).

However, literatures on GDP forecast are vast, starting to state a great work of Xavier Sala-i-martin (1997), who performs a modeling approach that depends on a million regression and hundreds of independent variables subdivided in 9 categories, these categories refer to the covenant source. Findings are also countless because he did an excessive study for underling the causal covariates correlation to GDP as well.

Martin (2019) utilized quarterly South African GDP data to assess forecasting performance across two model categories: traditional econometric models (Autoregressive and VAR), and SL methods (EN, RF, SVM and RNN). By achieving lower RMSE and higher correlation with actual trends, this paper highlights the potential of EN model to substitute traditional models (AR and VAR) and advanced models (RF, SVM, and RNN) for forecasting South African GDP. SL offers policymakers a powerful tool to navigate macroeconomic complexities.

The findings of Martin (2019) align with Maccarrone et al. (2021) studies on U.S. GDP forecasting who aims to provide a comparison of time series, linear regression, and machine learning models. The study predicted the Quarterly U.S. GDP through Seasonal Auto-Regressive Integrated Moving Average with exogenous factors (SARIMAX), a conventional linear regression

model (LR), averaging forecasts of Mean LR-SARIMAX¹, KNN, AR and ARX. The study uses two prediction strategies; one-step ahead and multi-step ahead. The One Step Ahead Forecasting Strategy forwards the counts of forecasting towards one quarter. Whereas, multi-Step Ahead Forecasting Strategy are run through the prediction of multiple steps time series forecasting for each horizon. As a results; The study concluded that the machine learning via KNN has showed best performance for the one-step-ahead strategy, with respect to SARIMA and AR (which are overestimated). On the other hand, it loses predictive power when the forecast is performed for a longer horizon.

Similarly, Xie and Yu (2020) perform a forecasting study on the driver of GDP growth of Singapore using Survey of Professional Forecasters (SPF). Five methods are used for this purpose, whereby comparing to the sample median of SPF forecasts. The results show MSPF of hybrid method drop mean squared forecast error (MSFE) by about 50% over that of the sample median, verifying to be statistically significant at the 5% level.

Messaoudi(2024)initially investigated the efficiency of Random Forest (RF) method to predict GDP growth in Algeria during the period 1964 to 2022, employing a rigorous four-step methodology, encompassing data preprocessing, feature selection, model training, and result validation. Although, this study does not detail used approach, findings mentioned the ability of RF to capture complex economic data patterns, outperforming conventional VAR framework. Building upon this study, Messaoudi and Khouidmi(2024) conducted extended comparative paper for the same purpose, comparing RF and Gradient Boosting Machine (GBM) against the traditional Vector Autoregressive (VAR) model for key macroeconomic indicators (GDP, Money Supply, inflation) in Algeria. Like previous paper, their research showed RF outperforming both GBM and VAR in accuracy and reliability, underscoring the significant potential of advanced machine learning techniques for improving economic forecasting and informing policy in emerging economies.

Correspondingly, Sahedet al. (2020) have demonstrated the performance of the Adaptive Neuro-Fuzzy Inference System (ANFIS), another SL model, compared to the conventional ARIMA(1,1,1) timeseries model. This study did not give neither resulting model nor predictions, otherwise findings stated that the ANFIS(1) model provided the most accurate GDP forecasts for Algeria, based on the root mean squared error (RMSE) criterion.

Controversially, another study, performed by Jijo (2025), on GDP Growth of G7 countries (Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States) suggested that ARIMA models may frequently outperform models based on data for specific purposes, key among capturing the historical trends of GDP fluctuations.

2.4 Advantages and disadvantages

Nevertheless, recent researchers have issued warnings about using complex modeling techniques (based on data patterns) to make predictions, leading to phenomenon known as overfitting the data (James, Witten, Hastie, & Tibshirani, 2013) and fallacious conclusions. This happens because SL procedure is working too hard to find patterns in the training data, and may

¹ The mean average ARIMAX-LR model is obtained through the average of two models accordingly:

$$\hat{Y} = \frac{1}{2} \times \hat{Y}_{ARIMAX} + \frac{1}{2} \times \hat{Y}_{LR}$$

be picking up some patterns that are just caused by random chance rather than by true properties of the unknown function . in this case, the accuracy metrics is not optimal (such as Mean Square Error of validation sets is very large).

As a conclusion, while indicative learning has demonstrated a good forecasting accuracy in many occasions, conventional models often provide more logically consistent explanations aligned with former established economic frameworks. Practically, GDP forecasting is not founded by a singular economic model but rather requires complex and dynamic reinforcements that may not always align with conventional theories, especially those derived from capitalist economies (Mankiw, Romer, & Weil, 1992). These divergences likely intensify the advantages of integrating new adaptive models, offering resilience to economic fluctuations and structural inconsistencies volatility. particularly in capturing Algeria's complex economic dynamics. However, challenges related to data scarcity, model interpretability, and long-run dynamics highlight the need for further research. By addressing these gaps, future studies can provide policymakers with robust tools to navigate Algeria's unique economic challenges and foster sustainable growth.

2.5 hybrid approaches

The hybrid machine learning and econometrics was a subject of statistical learning. Studies by Zheng et al. (2017); Khan (2022); and others, demonstrating the superior accuracy of hybrid models. Zheng et al. (2017) suggested a hybrid model by integrating autoregressive distributed Lag (ARDL) with Random Forest (RF) algorithms to forecast GDP in oil-exporting countries, including Algeria. By combining ARDL's ability to model long-run cointegration with Random Forest's strength in capturing non-linear patterns, their approach reduced RMSE by approximately 15% compared to standalone ARDL models. However, the study's reliance on limited sectoral data constrained its applicability to non-hydrocarbon sectors, a critical area for Algeria's diversification efforts. Khan (2022) has developed this pattern by integrating Long Short-Term Memory (LSTM) neural networks into a vector autoregressive (VAR) framework, focusing on dynamic interactions among hydrocarbon revenues, government spending, and demographic trends. this model achieved a 20% improvement in correlation with actual GDP trends, attributed to LSTM's capacity to model temporal dependencies and oil price volatility.

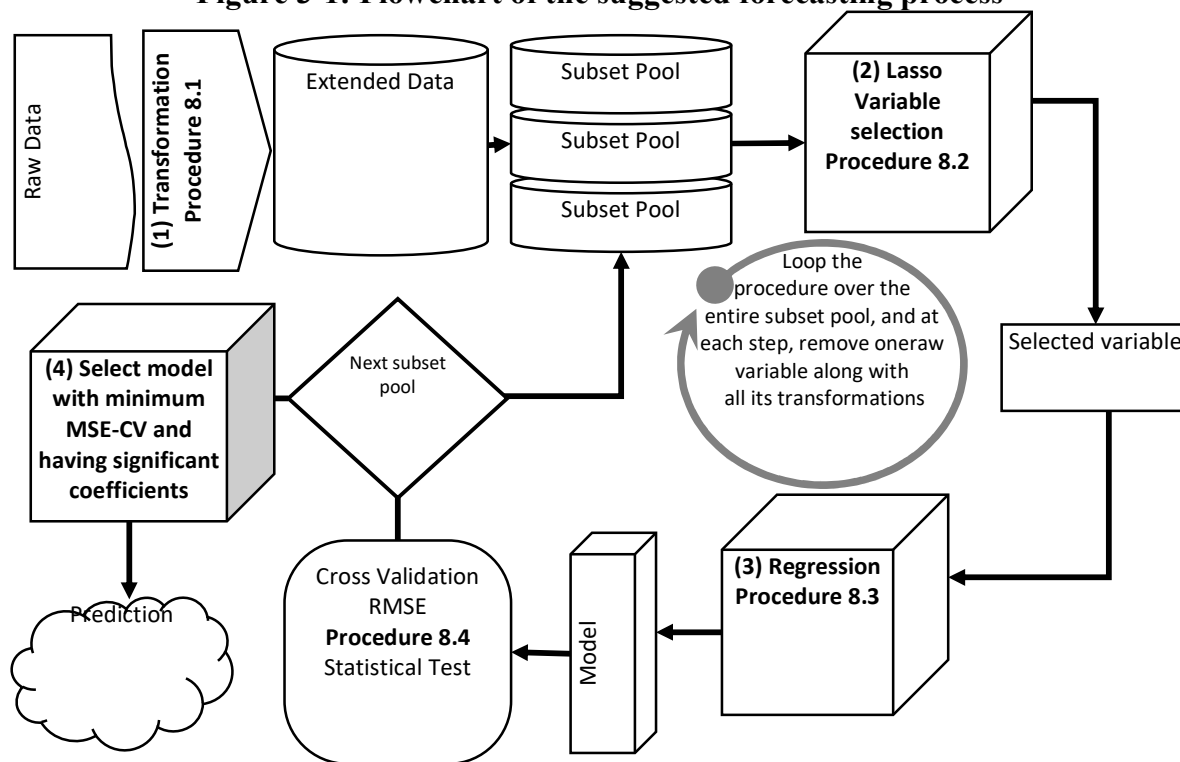
Although simple models like ARIMA, ARDL, and VAR are confessed to present a valuable insight over long time of use, but these frameworks have judged producing weak results with low accuracy (Messaoudi, 2024; Messaoudi & Khoudmi, 2024; Zheng, et al., 2017; Khan, et al., 2022; Maccarrone, Morelli, & Spadaccini, 2021; Martin, 2019), whereby data scarcity remained a challenge as frequent economic metrics of Algeria are often incomplete. Moreover, these models struggle to capture the nonlinear dynamics, structural breaks, and complex interactions of hydrocarbon-driven economy (Khan, et al., 2022), limiting their effectiveness in GDP forecasting in context of Algeria. Therefore, many studies have contributed to overcome these issues, and so on inspiring our research subject, shifting toward adopting more advanced and effective modern techniques, and integrating multiple approaches to achieve superior outcomes. Accordingly, this study aligned with hybrid approach, focusing on the integration of diverse methods through an

intelligent algorithm that performs various statistical tasks to develop an accurate and suitable model, considering optimal predictive fit within the context of the Algerian case study.

3. PROPOSED METHODOLOGY AND RESEARCH DESIGN

Inductive learning algorithms have two phases: learning and predicting (Hastie, Tibshirani, & Friedman, 2009). Meaningly, in this study we interested to supervised learning, firstly to learn which independent variable are mostly explain the dependent variables (Real GDP in our study), and secondly to employ these selected variables through regression models for predicting the outcome. The following figure steps the procedure to perform these purposes which is detailing thereafter.

Figure 3-1: Flowchart of the suggested forecasting process



Source: Authors compilation.

The Figure 3-1 outlines the proposed procedure, which consists of four major steps: data transformation, variable selection, model fitting, and model selection. Initially, the dataset undergoes transformation, including lagging and scaling, as part of the **Variables Processing (transformation)** step. This transformation extends the dataset to incorporate multiple patterns for each variable. The extended dataset is partitioned into multiple subsets, or folds. Each fold comprises the entire dataset except for one raw variable and all its associated transformations, which are systematically excluded. This exclusion facilitates the **Variable Selection** process in the subsequent step, enabling a balance between model complexity and predictive accuracy. The second and the third steps are executed iteratively. Subsequently, variable(s) selected in the second process is/are then used to specify the model (**Model Specification**), both with and without an intercept. Within these iterative steps, the resulting models are subjected to statistical

tests. After proceeding all folds, the final step involves applying selection criteria to models that meet parameters significance test and several model diagnostic tests. From these, the algorithm selects the best-performing model based on the minimum Mean Squared Error (MSE) achieved during the validation process. The following subsections provide a detailed explanation of each of these major steps.

3.1 Variables Processing (transformation)

Conventionally, most methods for estimating a function assume a linear form. In some situations, such an assumption is reasonable or even desirable. But often the true relationship is more complicated, in which case a linear model may not provide an accurate representation of the relationship between the input and output variables (James, Witten, Hastie, & Tibshirani, 2013). Given the inherent limitations of linear models, real-world data frequently exhibit intricate relationships often obscured by nonlinear patterns (Marwala, 2013). To address this issue, data scientists employ strategies such as variable transformations (e.g., logarithmic, exponential, or square root transformations) or incorporate smooth terms (e.g., splines or polynomial functions) (Hastie & Tibshirani, 1986) to achieve a more accurate representation of such relationships (Quintana, 2021). Mathematically, the newly transformed variable, denoted as $\tilde{x}_t$, can be expressed in ways depending on the nature of the transformation, inherent the following expression:

$$\tilde{x}_t = \phi(x_t)$$

where ϕ represents a transformation function applied to the original variable, or delayed effects such as:

$$\tilde{x}_t^j = x_{t-j}$$

where j denotes the lag order, capturing the effect of past values in time series data.

In this study, a series of preprocessing steps was implemented to systematically select appropriate patterns within the macroeconomic data in a relationship with the GDP, the data processing step includes:

- Feature transformation to capture nonlinear interactions;
- The application of temporal lag operators, enabling the identification of latent effects within time series data;
- Per capita setting shares aggregates for each person, scaling the variable to the total population.

This step aims to ensure capturing both nonlinear interactions in cross-sectional data and time-dependent dynamics as well, extending the dataset dimension and counting vast variable patterns. Afterward, the **Procedure 8-1** shows the R function code for preprocessing data. Subsequently, feature selection techniques introduced in next section filters the most relevant variables influencing GDP. This systematic approach allowed a more accurate analysis of macroeconomic patterns and their correlation with GDP trends.

3.2 Variable Selection

A hardest step of the model-building process and a very famous question of econometricians is; what predictor should be selected? (Ratner, 2017) For this reason, many

methods are developed, regularly categorized in conventional settings and new unseen settings. The conventional settings are based what economic theories have declared, in which predictors influence the target variable by definition.

When there are too many confused predictors, statisticians need calling for a variable selection method to choose potential regressors (Ratner, 2017). For instance, the GDP after the Keynesian theory is motivated by the economic policy (Mankiw, 2015), mainly the fiscal policy tools such as taxes and governments expenditure. Regarding unseen settings, predictors influence the target variable in more complicated manner, in which econometricians need their former experience and more advanced assessment based on data to support their suppositions. For instance, the GDP trend may be influenced by other variable such as the climate change or political circumstances as well as magnitude of the country in attracting foreign investment. This path is more realistic because it reflects the actual situations and could include indirect predictor and minimize the error.

Broadly speaking of this technique, variable selection is a prior step before performing a regression (Ratner, 2017), choosing the right variables (accounting conventional settings and/or indirect predictors) for building an accurate model. From stepwise to complete (all) variable selection, several techniques have been developed, thereby applicability depending on model complexity (Elliott, Gargano, & Timmermann, 2013) and data properties (Hastie, Tibshirani, & Friedman, 2009). However, Classical and modern variable selection methods for linear regression are categorized as subset selection, shrinkage, and dimension reduction (James, Witten, Hastie, & Tibshirani, 2013). Evenly, no unique method has been considered perfect for this purpose and no single method is proven optimal (Matloff, 2017).

While exhaustive approaches like complete subset regression (CSR) are resource-intensive and need best computer configuration (Khan, et al., 2022; Ratner, 2017) due to their need to assess all variable combinations. Other methods are based on Euclidean norms, or deem on dimension reduction (Matloff, 2017), which search the causal relationships between variable and closed distance between the target variable and candidate input variables (James, Witten, Hastie, & Tibshirani, 2013). Several methods are poorly producing performing models (Ratner, 2017), or not applicable when the number of variables (commonly denoted p) is much larger than the number of observations (commonly denoted n) (James, Witten, Hastie, & Tibshirani, 2013; Matloff, 2017).

By considering the limitation of the linear causation, real-world are more complicated whereas some relations are hidden beyond nonlinearity assumption (Marwala, 2013). Later data-scientist advocate a more efficient solution using regularization, refining by shrinkage methods (Hastie, Tibshirani, & Friedman, 2009). Among regularization options like LASSO, Ridge and the Elasticity Net (EN), these methods fit a large set of predictors but retaining predictors that only improve the prediction accuracy and enhance interpretability (James, Witten, Hastie, & Tibshirani, 2013). In this approach, we write an augmented error function (Alpaydin, 2014):

$$\text{Error} = \text{error on data} + \lambda \times \text{model complexity}$$

Given the linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:

- $\mathbf{X}$ is $(n \times p)$ the design matrix of features (with a column of ones for the intercept);
- n is the number of observations and p is the number of regressors;
- $\mathbf{y}$ is $(n \times 1)$ vector of target values;
- $\boldsymbol{\beta}$ is the coefficient vector;
- and $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is the error term.

Likewise, Least Absolute Shrinkage and Selection Operator regression (LASSO), also known as basis pursuit (Hastie, Tibshirani, & Friedman, 2009), is primarily a regularization technique, belonging to linear regression (Quintana, 2021). LASSO is founded on an augmented error function, so called cost function (Hastie, Tibshirani, & Friedman, 2009), which is defined by a Lagrangian form:

$$\min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \times \sum_{j=1}^p |\beta_j|$$

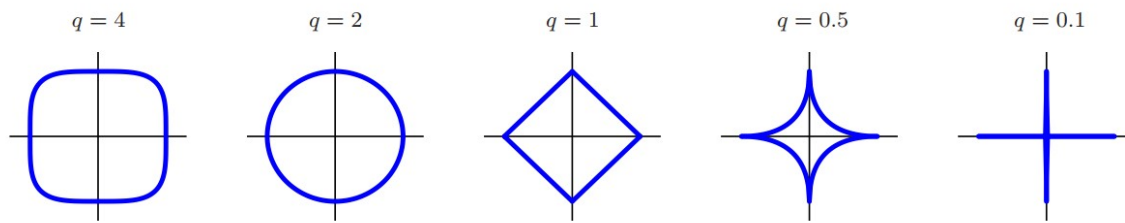
where:

- $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is residual sum of squares (RSS);
- $|\boldsymbol{\beta}|_1 = \sum_{j=1}^p |\beta_j|$ is L1 norm of the coefficient vector (penalty term which encourages sparsity), excluding the intercept in some implementations, but we'll include it here for simplicity;
- The L1 norm of a coefficient vector is given by $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$.
- λ : Regularization parameter (tuning hyperparameter).

Other LASSO Lagrangian form was deduced by Hastie et al. (2015), conducting to the same optimal $\boldsymbol{\beta}$, which is written as:

$$\min_{\boldsymbol{\beta}} = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

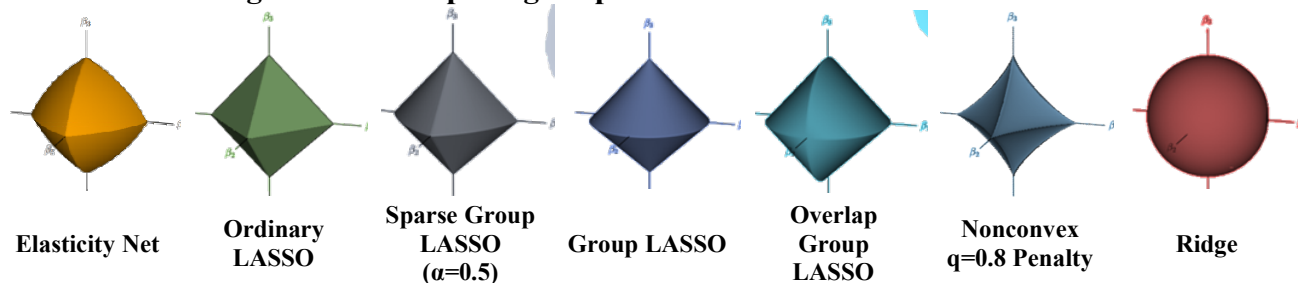
Where the index q in $\|\cdot\|_q$ denotes the q -th norm, key among indexing $q=2$ for the ridge, or equivalently $q=1$ (or L1 penalty term) for LASSO, and 0 (or L0 term) to count the number of non-zero coefficients (Hastie, Tibshirani, & Wainwright, 2015) which corresponds to variable subset selection (Hastie, Tibshirani, & Friedman, 2009). The **Figure 3-2** displays the constraint of regulation in case of two regressors ($p=2$).

Figure 3-2: the constraint region

Source: Hastie et al. (2015), page 22.

The LASSO regression extends ordinary least squares (OLS) by adding the L1 regularization term (penalty), shrinks useless coefficients (set inconsistent coefficients to zero) (Matloff, 2017) and bounds the size of predictions, thereby preventing overfitting (Quintana, 2021), mitigating multicollinearity (Tibshirani, 1996), and highlighting the most influential variables and then enabling variable selection (Tibshirani, 1996; Hastie, Tibshirani, & Friedman, 2009; Matloff, 2017). The LASSO is a well-known statistical learning technique for variable selection (Quintana, 2021), particularly in cases the number of features exceeds observations (James, Witten, Hastie, & Tibshirani, 2013). When the tuning parameter is sufficiently large, LASSO forces certain coefficient estimates to exactly zero, effectively performing subset selection (James, Witten, Hastie, & Tibshirani, 2013; Matloff, 2017).

However, the LASSO-derived models are typically more interpretable than formerly ridge regression, as LASSO tends to produce sparse estimator of (Matloff, 2017) by including only a subset of influential predictors (James, Witten, Hastie, & Tibshirani, 2013). Although these results do not ensure an optimal set of predictors (Matloff, 2017), LASSO was proved to be more appropriate for variable selection (Hastie & Tibshirani, 2020; Quintana, 2021; Zou, Hastie, & Tibshirani, 2006). Stating, many lasso versions were developed, key variants include Group LASSO, Elastic Net, Adaptive LASSO, Fused LASSO, Graphical LASSO, the Smoothly Clipped Absolute Deviation (SCAD) penalty, 1-norm Support Vector Machine (L1-SVM), and Dantzig Selector, but no of this method was proved better than ordinary (Standard) LASSO discussed in this study (Hastie, Tibshirani, & Wainwright, 2015). **Figure 3-3** below shows some sorts of lasso shapes.

Figure 3-3: Comparing shapes of LASSO and derived method

Source :Hastie et al. (2015), pages 58, 59,64, 67, 84, and 291.

Consequently, our study employs LASSO regression for its balance of simplicity and effectiveness in selecting a subset of influential variables. However, the initial phase of analysis acts as a filter to clarify causal links between target variable and predictors. **Figure 8-1** demonstrates the selection, case when $p=3$, the augmented error function shows that the LASSO estimates minimize the residual sum of squares (RSS) among all points within the rhomboid (diamond) shaped constraint $|\beta_1| + |\beta_2| + |\beta_3|$.

Based on the mathematical framework, Efron et al.(2004) developed an algorithm for solving LASSO, called Least Angle Regression (LARS). Confronted with the quadratic method, this algorithm computes the efficient LASSO solution path, providing coefficients for all possible values of the regularization parameter (Hastie, Tibshirani, & Wainwright, 2015). In this study, we do not derive its algorithm; instead, we demonstrate a straightforward implementation of the LARS algorithm for LASSO in R, as described in **Procedure 8-2**.

3.3 Model Specification

However, defining function form can be conducted by regression, which varies from variable selection, typically specifies variables of a model, in reflecting their understanding or theory of causal relationships (Hastie, Tibshirani, & Friedman, 2009). The current step focuses on generating a single-equation. Thereby statistical models normally specify how one set of variables, called dependent variables (or response variable which is generally denoted Y), functionally depend on another variable or set of variables (Charpentier, Flachaire, & Ly, 2018), called independent variables (or predictors which is generally denoted $(X_1, \dots)$ (Matloff, 2017).

Conventionally, the relationship between the dependent variable and independent variable is generally written as:

$$Y_t = f(X) + \varepsilon_t$$

is unknown function of $X_1, \dots$, and ε_t is a random error term, which is independent of X with zero mean and regular deviation (James, Witten, Hastie, & Tibshirani, 2013), assume that $\hat{Y}_t$ denotes the information provided by X_t to predict Y_t (Efromovich, 1999). The main reasons to estimate $\hat{Y}_t$ are either prediction and inference (James, Witten, Hastie, & Tibshirani, 2013). Prediction involves the situations in which the inputs X_t are already known and an output Y_t would be obtained. In this situation, the error term average tends to zero, predicted $\hat{Y}_t$, which is donated by $\hat{f}(X)$, is defined by the following estimation function:

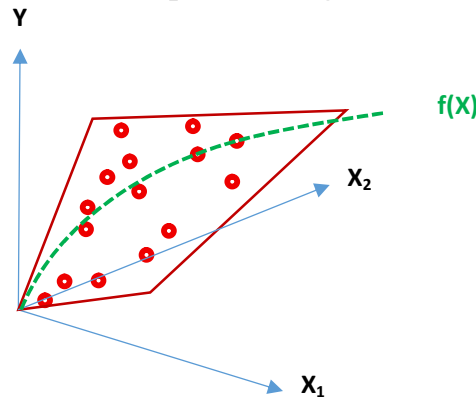
$$\hat{Y}_t = \hat{f}(X) = Y_t - \varepsilon_t$$

is a box system, which may provide a model that yields accurate predictions for Y :

$$X_t \rightarrow f \rightarrow \hat{Y}_t$$

Figure 3-4 illustrates how the function fits the predictions in case of two input variables ($p=2$).

Figure 3-4: 3D plot of a regression function



Source: Authors compilation.

The figure above plot visually a regression line which function $y = f(X)$. The two independent variables X_1 and X_2 , span the horizontal planes X-axis, while the dependent variable y , is showed along the vertical Y-axis. The contoured red surface illustrates how y varies as a function of X_1 and X_2 , providing a graphical representation of the model's predictions across the input space. Consequently, any point of y is predicted given the corresponding coordinates of X_1 and X_2 , allowing for estimations based on the available data points.

In the third step of our modeling approach, we define the functional form, denoted as f , for the target variable (Real GDP). Assuming a linear functional form, this step is implemented using an R command as specified in **Procedure 8-3**, altering this regression step to compound models with and without intercept. The input variables are those associated with non-zero coefficients derived from the LASSO-LARS algorithm, as outlined in prior **Procedure 8-2**.

3.4 Measuring the Quality of Fit

In statistical learning, the most prediction techniques are really interesting to find how the function f fits the unseen observation used in testing the response, rather how response function f fits in the training data (Xie & Yu, 2020). Accordingly, for measuring the quality of fit of an arbitrary function, several criteria were developed, similarly defining the residual sum of squares (RSS) by:

$$RSS(f) = \sum e_i^2$$

or equivalently as:

$$RSS(f) = \sum (y_i - f(x_i))^2$$

Minimizing R leads to infinitely many solutions: any function f passing through the training points (x_i, y_i) is a solution. Any particular solution chosen might be a poor predictor at test points different from the training points. If there are multiple observation pairs (x_i, y_i) each value of y_i , the risk is limited. In this case, the solutions pass through the average values of the x_i at each y_i (Hastie, Tibshirani, & Friedman, 2009).

Unlike The mean squared error (MSE) which determine how the expected response value may tend to the real response value for these observations. a MSE of cross-validation criterion (MSE-CV) is supposed to lead accurately to this opportunity (Hastie, Tibshirani, & Friedman, 2009; Matloff, 2017), finding the appreciated function and restricting the overfitting contradictory. Instead of using all collected data, CV techniques split the data into two parts, one is called training set (Matloff, 2017), the rest is called validation set or hold-out set (Hastie, Tibshirani, & Friedman, 2009). The training set is the data to conduct a regression, fitting the estimates. Based on the test set, the fitted model is used then for predicting values of the target variable, pretending its response variable values are unknown (Matloff, 2017).

The resulting validation set error rate provides an unbiased estimate of the true test error rate. Specifically, the Mean Squared Error (MSE), when applied in a cross-validation context (MSE-CV), quantifies the average squared difference between predicted values and actual outcomes for new, unseen data, thereby serving as a crucial metric for gauging a model's accuracy and generalization ability. While informal or ad-hoc validation techniques can be ambiguous, well-established methods such as Leave-One-Out Cross-Validation (LOOCV) and K-Fold Cross-Validation provide systematic and robust approaches for comprehensive model evaluation. For optimal model selection, the parameters derived from the training sets must be rigorously validated by assessing their predictive performance on the testing set. This process is essential for evaluating the model's ability to generalize to future, unseen values of the response variable, thus yielding the most reliable cross-validation criterion.

Consequently, the K-fold cross-validation algorithm is employed to ensure an unbiased data sample for model training. This is achieved by randomly permuting the input variable values and then dividing the observations into k equally sized folds. The algorithm iteratively calculates the Mean Squared Error (MSE) for each fold. The overall cross-validated MSE is then computed as the average of the MSE values obtained from iteration, as demonstrated by the following formula:

$$RMSE - CV = \sqrt{\frac{1}{k} \sum_{j=1}^k MSE_j}$$

The **Procedure 8-4** illustrates the R function used to calculate this criterion, assuming fivefoldssetting (k=5).

According to this formalization, the algorithm seeks to conduct a well specified model, thereby tending to select the final model meeting minimum RMSE-CV criterion and having significant feature(s). The following section provides empirical application of the proposed approach.

4. EMPIRICAL RESULTS

This section examines the implication of the proposed methodology for analysing Algeria's macroeconomic performance, developing a model to predict GDP, and deriving informed economic policy recommendations.

4.1 Research design and methodology

The empirical analysis builds on integrating supervised statistical learning techniques and ordinary regression, discussed in the previous section. We implemented this framework using the R programming language. The training dataset, sourced from Algeria's macroeconomic aggregates, spans the period from 1990 to 2023. This yearly dataset was extended through transformations to capture nonlinear patterns and shifted back to seize the delayed effect of economic policy, enabling model fitting and assessment of the impact of government policies on economic capacity.

4.2 Data

Consequently, we collect a data of this paper for Algeria economy to train the algorithm from a verity of sources, referring to Finance Act for each concerned year, The Bank of Algeria Reports, and the World Bank Open Data. Knowing that the algorithm renders this data to some transformation (Logarithmic, Polynomial, Rational, Progression, and lagged data), for intelligently capturing data patterns.

The dataset employed in this study includes monetary policy aggregates supplied by the Central Bank of Algeria. Additionally, fiscal policy and taxation data were obtained from the Finance Act issued by the Ministry of Finance of Algeria. The dataset encompasses macroeconomic aggregates and economic indicators, categorized as follows:

- Demographic metric counting total population;
- Economic policies involving monetary aggregates, government expenditures, taxation, and trade balance;
- Indicators of economic stability including Inflation, Unemployment, and Reserve of Gold and Foreign Currencies.

For further details, raw data are referenced to GitHub link on section **Declarations and Data Availability**.

4.3 Estimation Results

Within an initial pool of an extended dataset encompassing 1,062 raw variables and their corresponding transformations, as outlined in Erreur ! Source du renvoi introuvable. for data preprocessing, 58,410 candidate regression models were generated. Variable selection was subsequently conducted using the Least Angle Regression (LARS) algorithm, in accordance with the methodology specified in Erreur ! Source du renvoi introuvable.. Models counting statistically significant parameters, as estimated via the approach detailed in Erreur ! Source du renvoi introuvable., were retained for the model selection phase. The final model was selected based on the minimization of the Root Mean Squared Error under Cross-Validation (RMSE-CV), as described in Erreur ! Source du renvoi introuvable..

Table 4-1 reports the accuracy metrics associated with the selected model.

Table 4-1: Model Accuracy and Ranking Amongst the Selective Regressions

Metric	Value	Rank	Metric	Value	Rank
5-Fold MSE-CV	$3.250898 \times 10^{+11}$	1	MSE	$1.804853 \times 10^{+22}$	1
Ons-half-fold MSE-CV	$5.157898 \times 10^{+11}$	11	RMSE	$1.343448 \times 10^{+11}$	1

Adjusted R-Squared	0.99990688	1	R-Squared	0.99994541	1
Mean Absolute Error (MAE)	1.096064x10 ⁺¹¹	1	Efron's R-squares	0.93004663	27
Mean Absolute Percentage Error	0.00629851	1	Min-Max Accuracy	0.9937332	1
Akaike info criterion	54.91281	1	Schwarz criterion	55.47859	1

Source: Authors compilation.

This exhaustive process revealed the best model, based on lowest RMSE-CV criterion, characterised by metrics in **Table 4-2**.

Table 4-2: Model Characteristics

R-squared	0.999054	R-squared (no intercept formula)	0.99994541
Adjusted R-squared	0.998441	Adjusted R-squared (no intercept formula)	0.99990688
Standard Error of regression	1.75x10 ⁺¹¹	Akaike info criterion	54.91281
Sum squared resid (ESS)	5.23x10 ⁺²³	Schwarz criterion	55.47859
Mean dependent var	1.77x10 ⁺¹³	Hannan-Quinn criterion	55.09001
Durbin-Watson stat	2.233629	Log likelihood	-784.2358

Source: Authors calculations using R commands.

Except the Efron's R-squares and one-half cross validation MSE (2-Fold MSE-CV), the table above indicates that the model is the bestfitting databecause it is ranked optimum for several metrics, revealing the highest predictive accuracy of 99.99% (R-Squared is over 0.9999). So far comparing visually the actual GDP against predicted Real GDP, **Figure 8-2** shows consistent of the actual curve of Real GDP to the predicted Real GDP curve, supporting the predictive accuracy of minimal cross-validated Root Mean Square Error (RMSE-CV).

Moreover, using a large dataset significantly improved the variable selection procedure, particularly for forecasting Real Gross Domestic Product (RGDP) over a one-year horizon. The Mean Absolute Percentage Error (MAPE) of approximately 1.2% indicates a highly accurate forecast for $\square\square$ in 2024, with a predicted value of approximately 25.2 trillion DZD (see **Figure 8-2**) compared to the actual RGDP of 25.5 trillion DZD, as reported by the World Bank (2025).

Given these findings, model is conducted on 29 degrees of freedom, a simplified representation can be expressed as:

$$\begin{aligned}
 RGDP_t = & \beta_1 \times Unemployment_{t-1} + \beta_2 \times \log(RealExRate_t) + \beta_3 \times \log(RealExRate_{t-1}) \\
 & + \beta_4 \times \log(Import_t) + \beta_5 \times \log(M2_{t-4}) + \beta_6 \times \log(M2_{t-5}) + \beta_7 \times \frac{M2_{t-3}}{Pop_{t-3}} \\
 & + \beta_8 \times \frac{M2_{t-4}}{Pop_{t-4}} + \beta_9 \times \log\left(\frac{M2_{t-4}}{Pop_{t-4}}\right) + \beta_{10} \times \log\left(\frac{M2_{t-5}}{Pop_{t-5}}\right) + \beta_{11} \times \frac{FunExp_{t-3}}{Pop_{t-3}} \\
 & + \beta_{12} \times \frac{TotExp_{t-4}}{Pop_{t-4}} + \epsilon_t
 \end{aligned}$$

Where:

- are the regression coefficients;
- is the error term at time t;
- $\log$ is the natural logarithm of the covenant aggregate variable ;

- is the Per capita of the covenant aggregate , setting the share aggregates for each person;
- is the j^{th} lag period.

This model combines complex patterns of 8 raw variables, including:

- The Real Gross Domestic Product (**R**) as target variable;
- Amount Money supply ();
- Unemployment Rate (**Unemploy**);
- Real Exchange Rate from United States Dollars (USD) to local dinar (DZD) currency (**RealEx**)
- Total Imports in USD sourced from Bank of Algeria (**In**);
- Operating Budget (Functional Expenditures) (**Fu**);
- Total Expenditures (**To**);
- Total population () for capturing the share of aggregates for every person (Per Capita).

Consequently, the model is conducted through verity of transformation; counting:

- Logarithmic conversion of **M2** lagged by four and five years respectively;
- The share of money amount **M2** for every single person lagged by three and four years respectively;
- The Logarithmic conversion of the share of money amount **M2** for every single person lagged by four and five years at once;
- The logarithmic conversion of current real **Exchange rate** and one period lag;
- Lagged effect **unemployment rate** by one year;
- The logarithmic of the **total amount of Imports** valorised by USD;
- **Operating Budget** Per Capita lagged by three years;
- **Total Expenditure** Per Capita lagged by four years.

These transformations address potential non-linear relationships, while delayed variables render long-term dynamic impact on economic activity, acknowledging the changes may not immediately influence the economy but have been delayed to next period. Additionally, the Per Capita disposal reflects economic share for each person. Table 4-3 summarizes coefficients statistics.

Table 4-3: Coefficients statistics

	Estimate	Std. Error	t value	P-Value	
Unemployment_{t-1}	-4.437x10 ⁺¹⁰	1.981x10 ⁺¹⁰	-2.240	0.038774	*
Log(RealExRate_t)	1.689x10 ⁺¹²	7.819x10 ⁺¹¹	2.160	0.045352	*
Log(RealExRate_{t-1})	-3.018x10 ⁺¹²	8.777x10 ⁺¹¹	-3.438	0.003137	**
Log(Import_t)	1.249x10 ⁺¹²	2.643x10 ⁺¹¹	4.725	0.000196	***
Log(M2_{t-4})	-3.101x10 ⁺¹⁴	4.165x10 ⁺¹³	-7.444	9.59x10 ⁻⁰⁷	***
Log(M2_{t-5})	3.107x10 ⁺¹⁴	4.183x10 ⁺¹³	7.429	9.85x10 ⁻⁰⁷	***
M2_{t-4}/Pop_{t-3}	9.179x10 ⁺⁰⁶	3.146x10 ⁺⁰⁶	2.918	0.009593	**

M2_{t-4}/Pop_{t-4}	1.709x10 ⁺⁰⁷	3.399x10 ⁺⁰⁶	5.029	0.000103	***
Log(M2_{t-4}/Pop_{t-4})	3.081x10 ⁺¹⁴	4.125x10 ⁺¹³	7.468	9.19x10 ⁻⁰⁷	***
Log(M2_{t-5}/Pop_{t-5})	-3.103x10 ⁺¹⁴	4.194x10 ⁺¹³	-7.398	1.04x10 ⁻⁰⁶	***
FunExp_{t-3}/Pop_{t-3}	1.570x10 ⁺⁰⁷	5.830x10 ⁺⁰⁶	2.693	0.015400	*
TotExp_{t-4}/Pop_{t-4}	8.289x10 ⁺⁰⁶	2.455x10 ⁺⁰⁶	3.376	0.003587	**
Significance codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Source: Authors calculations using R command.

The above Table 4-3 shows that estimated coefficients, associated with each of the features, are significant which indicates rejecting the hypothesis of $H_0: \beta_j = 0$, thereby having less possibility to neglect the selected variables.

Further, although LASSO does not ensure regression assumptions, additional statistical tests are summarized in **Table 4-4**, reporting p-values of the major diagnosis statistics.

Table 4-4: Diagnostic Tests for Regression Assumptions

Model Validation Tests					
Durbin-Watson	2.2336	CHOW	0.2739	RESET	0.01
Jarque-Bera	0.7226	Skewness	2.322268	Kurtosis	-0.1399513
Homoscedasticity Test					
Breusch-Pagan	0.387	Glejser	0.497	Park Fitted	0.543
White	1	ARCH	0.7445	Harvey	0.6879
Autocorrelation Test					
Breusch-Godfrey	0.3365	Ljung-Box	0.7558	Pierce-Box	0.9138
Multicollinearity Test					
VIF1	VIF2	VIF3	VIF4	VIF5	VIF6
133.583	12464.200	15779.137	37779.3031	1370488819	1369569418
VIF7	VIF8	VIF9	VIF10	VIF11	VIF12
477.322	495.226	217963130	221025763	177.393	78.017

Source: Authors' compilation using multiple R commands.

The table above indicates that the model diagnosis tests respect several regression assumptions at a 5% significance level, with interpretations as follows:

- The Durbin-Watson statistic (DW = 2.2336) is close to 2, suggesting no significant first-order autocorrelation in the residuals.
- The Chow stability test (p = 0.2739) suggests no evidence of structural breaks, supporting parameter stability.
- The Ramsey RESET, Regression Equation Specification Error Test, (p=0.01) indicates potential model misspecification for one fitting terms, suggesting omitted variables or an incorrect functional form.
- The assumption of homoskedasticity is supported by several tests, including the Breusch-Pagan test (p = 0.387), Glejser test (p = 0.497), Park Fitted test (p = 0.543), White test (p = 1), ARCH test (p = 0.7445), and Harvey test (p = 0.6879), all with p-value > 0.05, indicating no significant heteroskedasticity.

- The absence of autocorrelation is supported by the Breusch-Godfrey test ($p = 0.3365$), Ljung-Box test ($p = 0.7558$), and Pierce-Box test ($p = 0.9138$), all with p -value > 0.05 , indicating no significant serial correlation in the residuals.
- The Jarque-Bera test ($p = 0.7226$) suggests that the residuals are normally distributed, supporting the normality assumption.
- Multicollinearity is present, as indicated by Variance Inflation Factor (VIF) values greater than 10 for several predictors. However, since the model is intended for prediction, this issue can be ignored (Gujarati, 2011).

Overall, the model appears robust, with no significant issues in autocorrelation, heteroskedasticity, or normality, as most tests yield p -values greater than 0.05. However, the Regression Equation Specification Error Test (RESET) yielded a p -value of 0.01, which is below the 5% significance level, requiring further investigation into other variable inclusion or functional form issue. But this test can be omitted ever DW statistics is close to 2, which indicates the model is well specified (Gujarati & Porter, 2009). However, economic models are only a serving tool to simplify and understand complex economic phenomena, though they are not without issues, as several studies have demonstrated. Moreover, no model could perfectly mirror reality, instead, they provide a framework for prediction (Ratner, 2017) and testing economic theories (Gujarati & Porter, 2009), including our study methodology. Therefore, what they have been encountered as issues, such as model misspecification and missing features, are a kind of model's limitations that statisticians and economists have long wrestled with (Cerreia-Viogliao, Hansen, Maccheronia, & Marinacci, 2022).

5. CONCLUSION

The contribution of this study aims to establish the macroeconomic model of Gross Domestic Products and specify the variable that influence the Real GDP, given macroeconomic data extracted of Algeria economy. Therefore, we extend classical approach of regression by a statistical machine learning approach, extending nonlinear patterns and reducing higher variable dimension by selecting most influential variables. In addition, using a large dataset improved the selection reliability for forecasting Real GDP ($RGDP_t$). Over a h -year horizon in the validation set, identifying the optimal model is based on the minimum MSE-CV criterion. This approach accurately achieved a high level of forecasting, establishing model with 99.99% of goodness of fit and reducing the overfitting as well.

Economically, driven-data models do not submit econometric assumptions and confronting theory interpretability. Controversially, the present conducted is founded on simple expression form, including different macroeconomic patterns and highlighting long-run dynamics. Consequently, lagged variables demonstrate delayed effect of economic policy through Money supply M2, Operating Expenses and Total Expenditures in the grow of the Real GDP. The unemployment rate plays a crucial role in this model, showing a negative relation with Real GDP. Total Imports valorised by US dollars and real exchange rates of US dollars to Algerian dinars express the effect of foreign trade on the fluctuation of the economic performance.

However, limitations of this study stand on comparing LASSO to other variable selection methods, revealing key distinctions. LASSO's inherent rigidity limits its ability to capture

complex relationships and interactions between variables, often leading to oversimplification. Furthermore, unlike methods such as CSR, which afford a more tailored modeling framework, LASSO's structure is less adaptable. Thereby further researches must integrate many variable selection strategies and address complex models' patterns. In addition, the non-linearity of several economic models involves future paper integrating fuzzy regression approaches.

Despite limitations in data availability, this paper emphasis for expanding future research to involve more countries case study. Furthermore, researchers to integrate new datasets into this framework for deeper insight, supported by initiatives focused on fostering data quality, refining forecasting accuracy, and improving economic performance.

6. DECLARATIONS AND DATA AVAILABILITY

- Competing interests: The authors declare that they have no conflict interest.
- The authors acknowledge their commitment to all publication conditions, including scientific integrity, respecting intellectual property rights, ensuring the article is free from plagiarism or duplication, additionally confirming that this research has not been previously published.
- Data raw can be downloaded via GitHub link:

<https://github.com/imedmail/Data/blob/501259b4f38fa696085ddb01b58ce4b9d0e6da7d/DataGDP.txt>

7. BIBLIOGRAPHY

- Abdelmounaim, R., & Abdelmounaim, H. (2024). Oil Price Shocks and Macroeconomic Performance of the Algerian Economy: A Structural VAR Approach. *مجلة بحوث الإدارة والاقتصاد*, 6(3), 536-544. <https://asjp.cerist.dz/en/article/255771>
- Alpaydin, E. (2014). *Introduction to Machine Learning* (3rd edition). The MIT Press.
- Barro, R. J., & Sala-i-Martin, X. (2004). *Economic Growth* (2nd edition). Massachusetts, The United State of America: The MIT Press.
- Binner, J. M., Elger, T., Nilsson, B., & Tepper, J. A. (2004). Tools for Non-Linear Time Series Forecasting in Economics: An Empirical Comparison of Regime Switching Vector Autoregressive Models and Recurrent Neural Networks. *Advances in Econometrics*, 19, 71–91. [https://doi.org/10.1016/s0731-9053\(04\)19003-8](https://doi.org/10.1016/s0731-9053(04)19003-8)
- Briggs, W. M. (2021, January). Models only say what they're told to say. In *International Econometric Conference of Vietnam* (pp. 35-42). Springer. https://doi.org/10.1007/978-3-030-77094-5_4
- Cerreia-Vioglio, S., Hansen, L.P., Maccheroni, F., & Marinacci, M. (2020). Making decisions under model misspecification. arXiv. <https://doi.org/10.48550/arXiv.2008.01071>
- Chakraborty, C., & Joseph, A. (2017). Machine learning at central banks (Working Paper No. 674). Bank of England. <https://www.bankofengland.co.uk/working-paper/2017/machine-learning-at-central-banks>
- Charpentier, A., Flachaire, E., & Ly, A. (2018). Econometrics and Machine Learning. *Economics and Statistics*, 505-506, 147–169. <https://doi.org/10.24187/ecostat.2018.505d.1970>
- Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer. <https://doi.org/10.1007/b97679>
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression (with discussion). *The Annals of Statistics*, 32(4), 407–499. <https://doi.org/10.1214/009053604000000067>

- Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, 177(2), 357-373. <https://doi.org/10.1016/j.jeconom.2013.04.017>
- Flachaire, E., Hacheme, G., Hué, S., & Laurent, S. (2022). GAM(L)A: An econometric model for interpretable Machine Learning. Communication Presented at French Stata Users' Group Meetings 2022 of the Conference No. 19, Aix-Marseille School of Economics. <https://doi.org/10.48550/ARXIV.2203.11691>
- Gujarati, D. N. (2011). *Econometrics by example*. Palgrave Macmillan.
- Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th edition). McGraw-Hill Education.
- Haponik, A. (2019, October 18). Machine learning in Economics – How is it Used? Retrieved November 18, 2022, from Addepto: <https://addepto.com/machine-learning-in-economics-how-is-it-used/>
- Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1(3), 297–310. <https://doi.org/10.1214/ss/1177013604>
- Hastie, T., & Tibshirani, R. (2020). Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4), 579–592. <https://doi.org/10.1214/19-STS733>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd edition). Springer. <https://doi.org/10.1007/bf02985802>
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.
- Hayes, A. (2022, May 12). *Econometrics: Definition, Models, and Methods*. Retrieved November 17, 2022, from Investopedia: <https://www.investopedia.com/terms/e/econometrics.asp>
- International Monetary Fund. (2024). Algeria: Selected issues (IMF Staff Country Reports). <https://www.imf.org/en/Publications/CR/Issues/2024/04/12/Algeria-Selected-Issues-547690>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.
- Jijo, J. J. (2025). Modelling and Forecasting the GDP of G7 Countries Using ARIMA Model. *Asian Journal of Economics, Business and Accounting*, 25(3), 295-315. <https://doi.org/10.9734/ajeaba/2025/v25i31711>
- Kalali, A., Richerson, S., Ouzunova, E., & Westp, R. (2019). Digital Biomarkers in Clinical Drug Development. In *Handbook of Behavioral Neurosciences* (Vol. 29, pp. 229-238). <https://doi.org/10.1016/B978-0-12-803161-2.00016-3>
- Khan, M. A., Abbas, K., Su'ud, M. M., Salameh, A. A., Alam, M. M., Aman, N., Mehreen, M., Jan, A., Hashim, N. A. A. B. N., & Aziz, R. C. (2022). Application of Machine Learning Algorithms for Sustainable Business Management Based on Macro-Economic Data: Supervised Learning Techniques Approach. *Sustainability*, 14(16), 9964. <https://doi.org/10.3390/su14169964>
- Kohlscheen, E. (2021). What does machine learning say about the drivers of inflation?. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3949352>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer Science.
- Liang, X. S. (2021). Normalized multivariate time series causality analysis and causal graph reconstruction. *Entropy*, 23(6), 679. <https://doi.org/10.3390/e23060679>
- Maccarrone, G., Morelli, G., & Spadaccini, S. (2021). GDP forecasting: Machine learning, linear or autoregression? *Frontiers in Artificial Intelligence*, 4, Article 757864. <https://doi.org/10.3389/frai.2021.757864>

- Madurapperuma, W. (2023). Money supply, inflation and economic growth of Sri Lanka: Co-integration and causality analysis. *Journal of Money and Business*, 3(2), 227–236. <https://doi.org/10.1108/JMB-08-2022-0039>
- Makhlouf, A. (2022). Measuring the Impact of The Non-oil Sectors on Gross Domestic Product in Algeria. *التكامل الاقتصادي*, 10(1), 418-433. <https://asjp.cerist.dz/en/article/187281>
- Mankiw, N. G. (2015). *Macroeconomics* (9th edition). New York, Ohio, United States of America: Worth Publishers.
- Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A Contribution to the Empirics of Economic Growth. *Quarterly Journal of Economics*, 107(2), 407-437. <https://doi.org/10.2307/2118477>
- Martin, L.-C. (2019). Machine learning vs traditional forecasting methods: An application to South African GDP (Working Paper No. 326). Department of Economics, Stellenbosch University. <https://ideas.repec.org/p/sza/wpaper/wpapers326.html>
- Matloff, N. (2017). *Statistical Regression and Classification: From Linear Models to Machine Learning*. CRC Press.
- Marwala, T. (2013). *Economic Modeling Using Artificial Intelligence Methods*. Springer.
- Messaoudi, M.(2024). Random Forest Modeling for Forecasting Economic Growth: A Machine Learning Approach in the Case of Algeria (1964-2022). *مجلة الأبعاد الاقتصادية*, 14(1), 391-410. <https://asjp.cerist.dz/en/article/249286>
- Messaoudi, M.,& Khoudmi, H. (2024). Comparative Analysis of Machine Learning and Autoregressive Models for Forecasting Economic Growth: A Case Study. *International Journal of Sustainable Development and Planning*, 19(8), 3049-3061. <https://doi.org/10.18280/ijstdp.190820>
- Mill, J. S. (2007). On the definition and method of political economy. In D. M. Hausman (Ed.), *The philosophy of economics: An anthology* (3rd edition, pp. 41–58). Cambridge University Press. <https://doi.org/10.1017/CBO9780511819025>
- Ouliaris, S. (2011, December). What is econometrics?. *Finance & Development*, 48(4), 2-10. <https://www.imf.org/external/pubs/ft/fandd/2011/12/basics.htm>
- Patanjali, K. (2018). *Machine Learning for Decision Makers: Cognitive Computing Fundamentals for Better Decision Making*. Apress.
- Quintana, R. (2021). Who Belongs in School? Using Statistical Learning Techniques to Identify Linear, Nonlinear and Interactive Effects. *The Quantitative Methods for Psychology*, 17(3), 312–328. <https://doi.org/10.20982/tqmp.17.3.p312>
- Ratner, B. (2017). *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data* (3rd edition). CRC Press.
- Reiss, P. C., & Wolak, F. A. (2007). Structural Econometric Modeling: Rationales and Examples from Industrial Organization. *The Handbook of Econometric*, 6, Part A, 4277-4415. [https://doi.org/10.1016/S1573-4412\(07\)06064-3](https://doi.org/10.1016/S1573-4412(07)06064-3)
- Sahed, A., Kahoui, H., & Mekidiche, M. (2020). Forecasting Algerian GDP using Adaptive Neuro Fuzzy Inference System During the Period 1990-2019. *Journal of Smart Economic Growth*, 5(2), 11-21. <https://ideas.repec.org/a/seg/012016/v5y2020i2p11-21.html>
- Sakiru, S., Gil-Alana, L., & Gonzalez-Blanch, M. (2022). Persistence of economic complexity in OECD countries. *Physica A: Statistical Mechanics and its Applications*, 603, Article 127860. <https://doi.org/10.1016/j.physa.2022.127860>
- Sala-i-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, 87(2), 178-183. <https://doi.org/10.2139/ssrn.31213>

- Sammur, C., & Webb, G. I. (2017). Inductive learning. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (2nd Edition, p. 1341). Springer. <https://doi.org/10.1007/978-1-4899-7687-1>
- Smaili, N., & Khelassi, A. (2021). The Relationship Between Public Expenditures and the GDP-growth in Algeria: An Empirical Study using the ARDL Model. *التكامل الاقتصادي*, 9(4), 847-861. <https://asjp.cerist.dz/en/article/179580>
- Stigum, B. (2003). *Econometrics and the Philosophy of Economics: Theory-Data Confrontations in Economics*. Princeton University Press.
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3-28. <https://doi.org/10.1257/jep.28.2.3>
- Vuppalapati, C. (2021). *Machine Learning and Artificial Intelligence for Agricultural Economics: Prognostic Data Analytics to Serve Small Scale Farmers Worldwide*. Springer.
- World Bank. (2025). *Algeria Economic Update, Spring 2025: Accelerating Productivity Gains for Diversified and Resilient Growth*. World Bank. <https://doi.org/10.1596/43358>
- Xie, T., & Yu, J. (2020). Forecasting Singapore GDP using SPF data (Working Paper No. 2020_017). Singapore Management University. https://ideas.repec.org/p/ris/smuesw/2020_017.html
- Zermane, H. (2025). Machine Learning Techniques for GDP Per Capita Prediction in Developing Countries: A Case Study of Sub-Saharan Africa. *Smart Computing and Control Renewable Energy Systems*, 516–524. https://doi.org/10.1007/978-3-031-80301-7_56
- Zheng, E., Tan, Y., Goes, P., Chellappa, R., Wu, D., Shaw, M., Sheng, O., & Gupta, A. (2017). When Econometrics Meets Machine Learning. *Data and Information Management*, 1(2), 75-83. <https://doi.org/10.1515/dim-2017-0012>
- Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2), 265–286. <https://doi.org/10.1198/106186006X113430>

8. APPENDICES

Procedure 8-1: R Function for Data preprocessing

```
Fun_TrfdData <- function(RData, Trfs) {
  base <- as.numeric(gsub("[^0-9]+", "", Trfs))
  fn <- gsub("[^a-zA-Z]+", "", Trfs)
  LRData1 <- Fun_LagData(RData, 1)
  if(fn == "" && !is.na(base)) return(RData^base)
  if(fn == "lag" && !is.na(base)) return(Fun_LagData(RData, base))
  switch(Trfs,
    "abs" = abs(RData),
    "sqrt" = sqrt(abs(RData)),
    "log" = log(abs(RData)),
    "ratio" = RData / LRData1,
    "prog" = (RData / LRData1 - 1) * 100,
    "PerCap" = RData / RawData[, "pop"],
    RData
  )
}
```

Procedure 8-2: R Function for variable selection using LARS algorithm

```
library('lars')
Fun_LARSelection<- function(y, X) {
  CData <- na.omit(c(y,X))
  y <- CData[,1]
  X <- CData[, -1]
  ys<- scale(y, center = TRUE, scale = FALSE)
  Xs<- scale(X)
  lars_model <- lars(Xs, ys, type = "lasso", normalize = FALSE)
  coef_lars <- coef(lars_model, s = 0.5, mode = "fraction")
  return(as.data.frame(X[,which(coef_lars != 0)]))
}
```

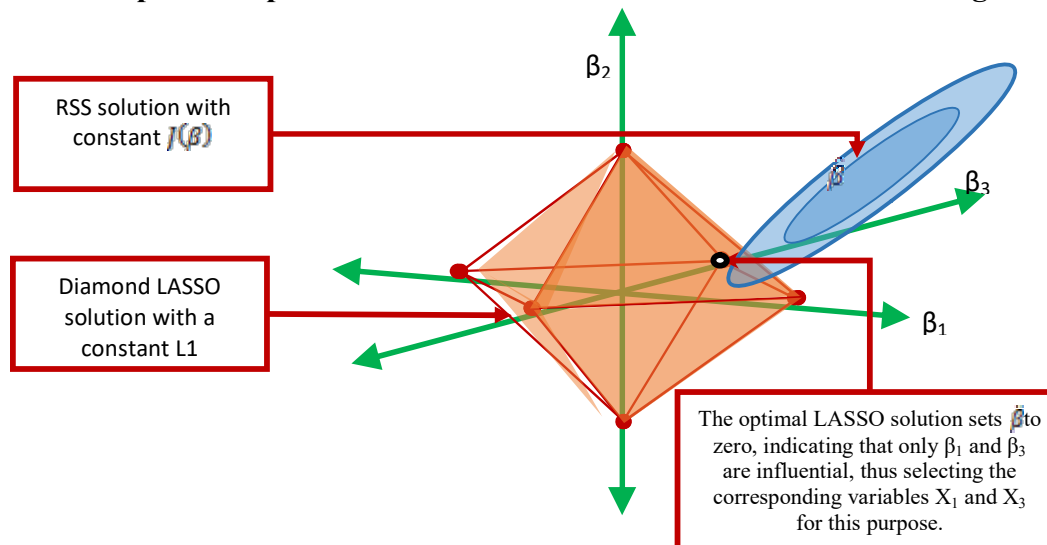
Procedure 8-3: R Command for Performing Model Specifications

```
#Fit the model with intercept
Fit_Model1<- lm(formula = RGDP.LC ~ ., data = Fun_LARSelection(RGDP.LC, X))
#Fit the model without intercept
Fit_Model2<- lm(formula = RGDP.LC ~ -.1, data = Fun_LARSelection(RGDP.LC, X))
```

Procedure 8-4: R Function to calculate Cross Validated Mean Square Error

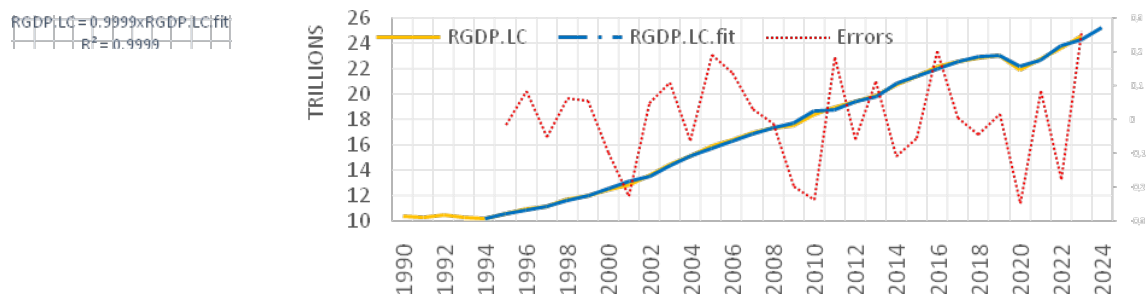
```
Fun_KFold_cv_mse <- function(Y, X, intercept, k = 5){
  set.seed(Sys.time())
  indices <- sample(length(Y))
  folds <- cut(seq_along(indices), breaks = k, labels = FALSE)
  mean(sapply(1:k, function(i){
    test_indices <- indices[folds == i]; train_indices <- indices[folds != i]
    Y_train <- Y[train_indices]
    Y_test <- Y[test_indices]
    X_train <- X[, train_indices, drop = FALSE]
    X_test <- X[, test_indices, drop = FALSE]
    model_fit <- lm(Y_train~X_train)
    (Y_test - predict(model_fit, newdata = X_test))^2
  })))
}
```

Figure 8-1: Graphical Aspect of LASSO for variable selection in case of 3 regressor ($p=3$)



Source: Authors compilation.

Figure 8-2: Actual Real GDP vs Predicted Real GDP



Source: Authors Compilation.