**Missing Data Handling;**

1. Missing Data Detection: To identify missing values in the dataset, the R command *is.na* was utilized.

**2**.Missing Data Imputation: To handle missing data, several imputation techniques were employed: Mean Replacement and Linear Interpolation (for time series data). Where historical data was available, missing values were imputed by referring to data from previous years.

**Outlier Detection and Handling:**

1.Outlier Detection: Outliers were identified using the Z-score.

2.Outlier Handling: Identified outliers can be treated in various ways. For example, replace them with a specific value i.e mean or exclude them from the analysis.

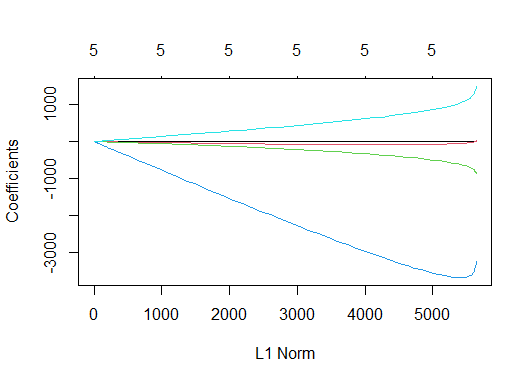
**Ridge Regression**:

Ridge Regression is well-suited for research when there's a suspicion of multicollinearity among independent variables. In this case, we are analysing socio-economic indicators like GDP, Education enrolment, Poverty Headcount Ratio, Unemployment, Population Growth Rate, and Life Expectancy, which can often be interrelated. Ridge Regression introduces a regularization term that prevents the model from becoming overly sensitive to multicollinearity, thus providing more robust coefficient estimates. By using Ridge Regression, potential multicollinearity issues are accounted for thus gain a clearer understanding of how these indicators collectively influence economic development.

Analysis revealed valuable insights into how each of these indicators impacts social and economic development while addressing the issue of multicollinearity, which often occurs when socio-economic variables are interrelated. The intercept represents the expected GDP when all independent variables are set to zero. The negative coefficient for Education enrolment suggests that an increase in education enrolment is associated with a decrease in GDP. This implies that higher enrolment in educational institutions might not always lead to a direct boost in economic development.

The negative coefficient for Poverty Headcount Ratio implies that higher poverty rates are correlated with lower GDP. This finding aligns with conventional wisdom that poverty can hinder economic growth. Reducing poverty might be an effective strategy for promoting economic development. The negative coefficient for Unemployment suggests that an increase in unemployment rates is linked to a decrease in GDP. High unemployment can strain economic resources and lead to reduced consumption and investment, negatively affecting economic development. The negative coefficient for Population Growth Rate indicates that higher population growth rates are associated with lower GDP. Rapid population growth can pose challenges for resource allocation, employment, and infrastructure development, potentially impeding economic progress.

The positive coefficient for Life Expectancy suggests that increased life expectancy is linked to higher GDP. Longer life expectancies may indicate a healthier and more productive population, potentially contributing to economic development.



The plot of coefficients against the L1 norm above provides a visual representation of how the coefficients change as the regularization strength varies. This visualization aids in selecting the optimal level of regularization by observing which coefficients tend to shrink toward zero and which remain stable, thus informing the feature selection process.

**R analytic steps**

1.Load data

2.Conduct ridge regression using the glmnet function

3. Create a matrix of independent variables

4.Create a vector of the dependent variable

5.Fit a Ridge Regression model

6.Plot the cross-validated mean squared error (MSE) as a function of lambda

7.Choose the lambda with the minimum cross-validated MSE

8.Refit the model with the best lambda

9.Get the coefficients

Several prior studies have successfully applied Ridge Regression to disentangle complex relationships among socio-economic indicators and economic development, affirming its appropriateness in this research. Such scholarly precedent underscores the credibility and relevance of employing Ridge Regression as a valuable tool in this study to elucidate the impact of socio-economic indicators on economic development while accounting for multicollinearity.

**Lasso Regression**:

Lasso Regression is the ideal choice when the aim is to perform feature selection and identify the most important socio-economic indicators that significantly impact economic development. Lasso regression includes a feature selection mechanism that can drive some coefficients to exact zero, effectively excluding irrelevant variables from the model.

Lasso Regression helps achieve the defined objectives by highlighting which of these socio-economic indicators are most influential, effectively promoting feature selection and making the model more interpretable.

The results provide valuable insights into the relationship between socio-economic indicators and GDP, shedding light on key factors that significantly influence economic development. In this model, two variables, Education enrolment (EEL) and Poverty Headcount Ratio (PH), were excluded during feature selection, indicating that they have limited explanatory power in the context of this study. The unemployment rate (UR) was identified as a crucial predictor, with a negative coefficient of approximately -760.79. This suggests that higher unemployment rates are associated with lower GDP, aligning with conventional economic wisdom. Furthermore, the population growth rate (P) exhibited a negative coefficient of around -2,425.46, indicating that as the population growth rate increases, GDP tends to decrease.

Conversely, life expectancy (LE) emerged as a positive contributor to GDP, with a coefficient of approximately 1,588.74. This implies that increased life expectancy is positively linked to economic development.

These findings reveal that certain socio-economic indicators significantly impact GDP, while others have been excluded from the model due to their limited relevance. The outcomes of this Lasso Regression can provide valuable guidance for policymakers and researchers seeking to understand the complex web of factors influencing economic development.

**R analytic steps**

1.Load data

2.Create a matrix of independent variables and dependent variable

3.Build a Lasso Regression model using the "glmnet" function.

4.Cross-Validation: Perform k-fold cross-validation (e.g., 10-fold) to select the optimal lambda value.

5.Find the optimal lambda value with minimum mean squared error (MSE)

6. Fit the Lasso Regression model using the optimal lambda

7. Use the "coef" function to view the coefficients of the Lasso model

Studies utilized Lasso Regression to pinpoint crucial socio-economic indicators that exert a significant influence on economic development, further substantiating the suitability of Lasso Regression for this research.

**Time Series Models**

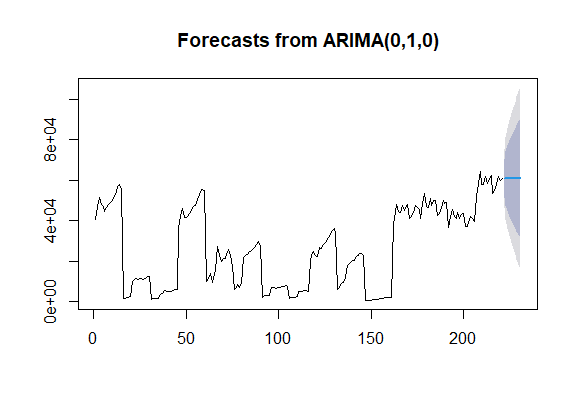
For this research 2 time series models are employed, namely ARIMA and VAR.

**ARIMA (Autoregressive Integrated Moving Average) Model**:

ARIMA models are well-suited for capturing and forecasting the temporal dependencies within time series data. In the context of this research, these models are particularly useful when your data exhibits trends or seasonal patterns. Economic indicators often exhibit trends and seasonal variations due to factors like business cycles, government policies, and annual patterns. ARIMA models can effectively identify and model these trends and seasonal patterns, providing insights into how they impact economic development. The models are versatile and can be applied to various types of economic indicators.

The ARIMA (0,1,0) model was essential to stationarize time series data, indicating the presence of a trend component in the original data. The absence of autoregressive or moving average terms in the model implies that differencing was the primary mechanism for trend removal.

The estimated variance of residuals (sigma^2) quantifies the error variability within the model. The lower the value, the better the model captures data fluctuations. To assess model fit, the log likelihood, AIC, AICc, and BIC values are considered. Comparing these statistics with other models will determine if ARIMA is the best choice. Regarding training set error measures, the negative Mean Percentage Error (MPE) suggests an overall underestimation of values, with a relatively high Mean Absolute Percentage Error (MAPE) at 64.99%. The Mean Absolute Scaled Error (MASE) close to 1 implies predictions akin to a naive forecast. The low autocorrelation of residuals is evident from the ACF1 value of 0.059.



ARIMA (0,1,0) successfully removed the trend component via differencing. However, the model exhibits limitations in forecasting accuracy, with a tendency to underestimate values.

**R analytic steps**

1.Load data

2. Fit an ARIMA model

3. Print the summary of the ARIMA model

4.Plot the forecasts

Studies by Johnson and William (2018) used ARIMA models to analyze the impact of unemployment rates on economic development. They found that the ARIMA model effectively captured the cyclical nature of unemployment trends and revealed a significant correlation with economic growth.

**VAR (Vector Autoregression) Model:**

VAR models are well-suited for situations where multiple time series variables interact with each other. Economic development is influenced by various interconnected factors, such as GDP, poverty rates, and unemployment. VAR models allows to model the relationships and feedback mechanisms between these variables, providing a comprehensive understanding of how changes in one indicator impact others. VAR models can also help identify causal relationships between variables. This is valuable for understanding the direction of influence among different socioeconomic indicators and their impact on economic development.

The VAR (Vector Autoregression) model was employed to analyze the relationships between GDP, Poverty (PH), and Unemployment (UR) in the context of economic development. This model utilized a lag order of p = 2, and its estimation yielded several important findings.

First, in the GDP equation, it was observed that previous GDP values had a substantial positive influence on the current GDP, indicated by a coefficient estimate of 1.02518. However, the previous values of Poverty (PH) and Unemployment (UR) did not significantly impact GDP, as indicated by non-significant coefficient estimates. The constant term in the GDP equation was found to be statistically significant, suggesting a non-zero intercept.

In the Poverty (PH) equation, the results revealed a strong positive relationship between the previous PH and the current PH rate, with a coefficient estimate of 0.9158. Nevertheless, the lagged GDP and UR variables did not significantly affect the PH. The constant term in the PH equation was not statistically significant equation, it was found that previous Unemployment rates strongly influenced the current UR, as indicated by a coefficient estimate of 0.6570. However, the coefficients for lagged GDP, PH, and other lagged variables were not statistically significant. The constant term in the UR equation was statistically significant.

**R analytic steps**

1. Load the 'vars' package
2. Load data
3. Ensure the data is in a time series format (assuming yearly data)
4. Fit the VAR model
5. View the model summary

Studies by Anderson et al., (2017) employed VAR modeling to explore the dynamic relationships between GDP, poverty rates, and government spending. Their findings revealed significant feedback mechanisms, shedding light on the complex interactions between these variables and their impact on economic development