Multiple linear regression report

Ahmad Jebril, December 10th 2024

Introduction:

The main goal of this project is to predict the GDP growth using a multiple linear regression model. The GDP growth is an important indicator of a country's economic performance, and accurate predictions are necessary for policy making, investment decisions, and economic planning. In this project, I utilized various economic indicators as predictors such as the consumption expenditure, investment, net exports, and government spending to model and predict the GDP growth. The analysis was done using the Global Economy Indicators dataset sourced from Kaggle.com, and the model's performance was then evaluated using the Root Mean Squared Error (RMSE) metric.

Multiple Linear Regression Methodologies:

Multiple linear regression is a statistical technique used to model the relationship between a dependent variable and multiple predictors, also known as the independent variables. In this project, I used multiple linear regression to predict the GDP growth (dependent variable) using several independent economic indicators in the Global Economy Indicators dataset, such as the consumption expenditure, investment, net exports, and the government spending (independent variables). The variables consumption expenditure and the government spending were combined and added together as the total consumption to reduce multicollinearity between the predictors.

The variable imports of goods and services was subtracted from the variable exports of goods and services to calculate the net exports.

The model was built using the least squares approach, which minimizes the sum of squared residuals between the observed and the predicted GDP values. The data cleaning process involved imputing the mean for certain columns and dropping null values for other columns handling missing values, and outliers were also addressed to ensure the model is robust. The data was split into training and testing sets before fitting the model, and the model's performance was evaluated using the Root Mean Squared Error (RMSE), which measures the average prediction error in the model.

Application:

1. Data Preprocessing:

   To prepare the data for analysis, the following steps were performed:

   - Missing values in multiple columns were imputed using the mean of the specific columns.
   - Using boxplots, outliers were visually examined and a decision was made to retain them in the model for the sake of maintaining the integrity of the data.
   - The data was split into 80% training data and 20% testing data. The model was trained using the training set, and the performance of the model was evaluated using the testing set.
   - The dataset was narrowed down to include only relevant variables (GDP, Total consumption, Investment, and Net Exports).

```
> summary(data_subset)
 Gross.Domestic.Product..GDP. Total_consumption   Gross.capital.formation  Net_Exports
 Min.   :2.585e+06            Min.   :2.544e+06   Min.   :-4.397e+10      Min.   :-8.600e+11
 1st Qu.:1.439e+09            1st Qu.:1.473e+09   1st Qu.: 2.838e+08      1st Qu.:-8.061e+08
 Median :8.071e+09            Median :7.810e+09   Median : 1.793e+09      Median :-9.564e+07
 Mean   :1.829e+11            Mean   :1.666e+11   Mean   : 4.654e+10      Mean   : 8.893e+08
 3rd Qu.:5.173e+10            3rd Qu.:4.632e+10   3rd Qu.: 1.284e+10      3rd Qu.: 1.848e+08
 Max.   :2.330e+13            Max.   :2.265e+13   Max.   : 7.600e+12      Max.   : 4.700e+11
```

2. Exploratory Data Analysis (EDA):

   A correlation matrix revealed a strong positive relationship between GDP and the predictors Total Consumption (0.993) and Gross Capital Formation (0.946), indicating an increase in these variables is closely associated with the GDP growth. However, a moderate negative correlation was observed between the GDP and the Net Exports (-0.439), indicating that higher net exports may result in a slight reduction in the GDP, suggesting the influence of economies relying on imports. Also, a strong positive correlation (0.903) between Total Consumption and Gross Capital Formation highlights the presence of multicollinearity, which was sorted during model evaluation.

```
> print(correlation_matrix)
                          Gross.Domestic.Product..GDP. Total_consumption Gross.capital.formation Net_Exports
Gross.Domestic.Product..GDP.                    1.0000000         0.9931737              0.9458005  -0.4389935
Total_consumption                               0.9931737         1.0000000              0.9026096  -0.5078446
Gross.capital.formation                         0.9458005         0.9026096              1.0000000  -0.2440160
Net_Exports                                    -0.4389935        -0.5078446             -0.2440160   1.0000000
```

3. Model Fitting:

   The multiple linear regression model was fitted to predict GDP using the predictors Total Consumption, Gross Capital Formation, and Net Exports. The model explained 99.7% of the variance in the GDP, as indicated by the R

squared value, and withall the predictors showing strong statistical significance ($p<0.001$). The coefficient for Total Consumption (0.840) indicates a strong positive relationship with GDP, while the Gross Capital Formation (0.921) and the Net Exports (0.464) also showing strong positive effects. The residual standard error is 1.721e+10 reflecting the typical prediction error and the overall F-statistic confirmed the model's significance ($p<2.2e-16$).

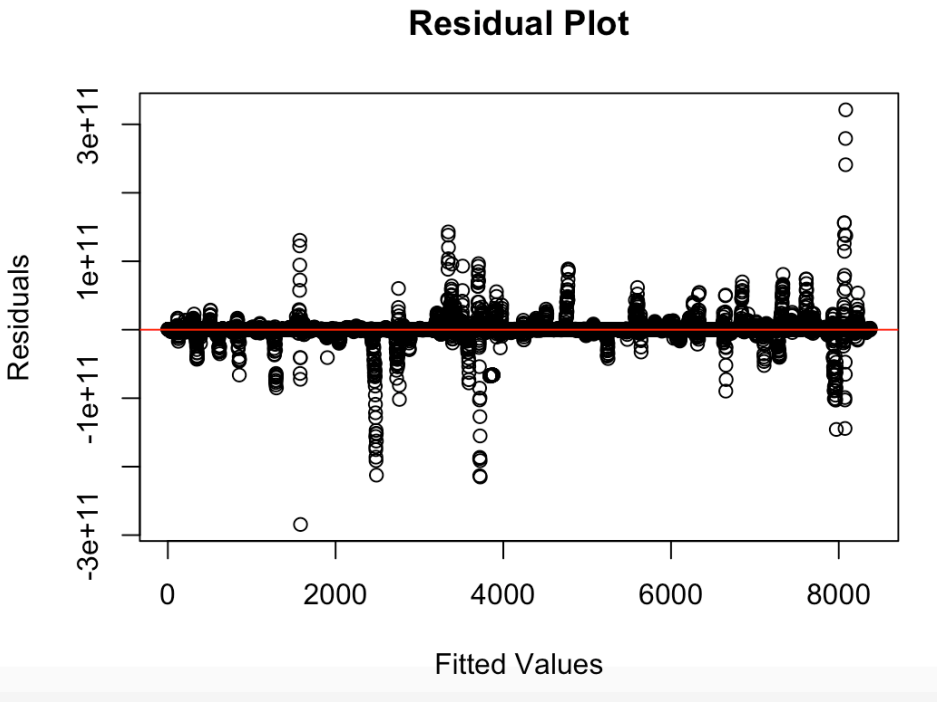
```
Call:
lm(formula = Gross.Domestic.Product..GDP. ~ Total_consumption +
    Gross.capital.formation + Net_Exports, data = train_data)

Residuals:
       Min         1Q     Median         3Q        Max
-2.844e+11 -1.918e+07  3.621e+08  7.799e+08  3.212e+11

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)             -3.772e+08  1.934e+08   -1.951   0.0511 .
Total_consumption        8.400e-01  6.871e-04 1222.614   <2e-16 ***
Gross.capital.formation  9.218e-01  2.073e-03  444.645   <2e-16 ***
Net_Exports              4.636e-01  8.013e-03   57.855   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.721e+10 on 8371 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9997
F-statistic: 8.348e+06 on 3 and 8371 DF,  p-value: < 2.2e-16
```

4. Model Evaluation:

   The model's performance was evaluated using the Root Mean Squared Error (RMSE) on the testing set, which was calculated to be 19.68 billion. This reflects the average prediction error and is reasonable given the scale of the GDP data, which spans trillions of dollars.

```
"RMSE: 19675224677.0176"
```

Residual analysis confirmed that the residuals are randomly distributed with no clear pattern observed in the residual plot, causing the assumption of homoscedasticity.



In addition, the variance inflation factor (VIF) for all the predictors was below 10, indicating acceptable multicollinearity and no severe interdependencies between the variables. Overall, the model demonstrates a strong predictive power and handles unseen data well, as seen by its test set performance.

```
vif(final_model)
    Total_consumption Gross.capital.formation             Net_Exports
             9.998948                8.111313                1.960570
```

Conclusion

In this project, a multiple linear regression model was built to predict the GDP using Total Consumption, Investment, and Net Exports as predictors. The model predicted 99.7% of the variance in GDP with an RMSE of 19.68 billion units indicating a strong predictive power. The model satisfies the assumptions of linear regression which was confirmed by residual analysis and multicollinearity was effectively managed. The results highlight the significance of Total Consumption and Investment influence on GDP growth.

References

1. Samiyu, M. (2021). *Multiple Regression Model for Predicting GDP Using Macroeconomic Variables*. SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3895177
2. Pennsylvania State University (n.d.). *Detecting Multicollinearity Using Variance Inflation Factors*. Available at: https://online.stat.psu.edu/stat462/node/180
3. Fang, Z. (2023). *Application of Linear Regression in GDP Forecasting*. ResearchGate. Available at: https://www.researchgate.net/publication/380849941_Application_of_Linear_Regression_in_GDP_Forecasting
4. Corporate Finance Institute (n.d.). *Variance Inflation Factor (VIF)*. Available at: https://corporatefinanceinstitute.com/resources/data-science/variance-inflation-factor-vif/
5. Kock, N., & Lynn, G. (2012). *Lateral Collinearity and Misleading Results in Variance-Based SEM: An Illustration and Recommendations*. Journal of the Association for Information Systems, 13(7), 546-580. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC6900425/