MULTIPLE REGRESSION MODEL FOR PREDICTING GDP USING MACROECONOMIC VARIABLES

(PART 1)

Mutiu Samiyu

Msamiyu2@gmail.com

July 28, 2021

Contents

Abstract

This research explores how one may predict the Gross Domestic Product (GDP) of a country using a technique known as multiple linear regression (MLR). Specifically, we explore whether other macroeconomic variables such as population, interest rates, unemployment rates, amongst others, can be used to predict the GDP of a country. We also examine the impact of new variables on the model base model fit using p-values and variance inflation factor (VIF) as a performance metric. The MLR model appears to be a suitable model for determining a linear relationship between dependent and independent features.

Introduction

This short essay explores how one may predict the Gross Domestic Product (GDP) of a country using a technique known as multiple linear regression. Specifically, we examine whether other macroeconomic variables such as population, interest rates, unemployment rates, amongst others, can be used to predict the GDP of a country. Linear regression is a statistical model for finding the relationship between input variables (also known as Independent/explanatory variables) and output variables (dependent/response variables). Linear regression models are used for predicting continuous variables compared to logistic regression models used for predicting categorical outcomes.

When we have just one input variable, it is a case of a simple linear regression whose equation is of the form;

$$Y = \beta_0 \; + \; \beta_1 X + \varepsilon$$

Where Y is the output variable, X is the input variable, β_0 is the intercept (i.e., the predicted value of Y when X is zero. β_1 is the slope/regression coefficient (i.e., the rate at which Y is expected to change when X increases/decreases. ε is the error term whose expectation is zero ($E(\varepsilon) = 0$(i.e., it is expected that the variation of the regression coefficient should be zero)

For multiple linear regression, which consists of more than one independent variable, the general equation for a set of observations with k linearly independent predictor variables is of the form;

$$Y_i = \beta_0 \; + \; \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

It should be noted that for $i = 1,2, \ldots n$ the relationship between the dependent variable Y and the k independent variables $x_{11}, x_{12}, \ldots, x_{nk}$ are linear in the parameters. The matrix form of the general linear model is written as; $Y_{N\times 1} = \beta_{k\times 1} X_{N\times k} + \varepsilon_{N\times 1}$

Parameter Estimation

The base model for linear regression is the “Least Square Estimate”. It squares and sums all the vertical deviations from each observation (data points) to the line, and the goal is to minimize the error. The least-square estimation is commonly used in fitting the regression line.

Normal Equations

The system of normal equation of the expression;

$$Y_i = \beta_0 \ + \ \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \text{ is given as;}$$

$$[y_1 \ \vdots \ y_n \ ] = [1 \, x_{11} \ \ldots \ \ x_{1k} \ \vdots \vdots \ddots \quad \vdots \ 1 \, x_{n1} \ \ldots \ \ x_{nk} \ ][\beta_0 \ \vdots \ \beta_n \ ][e_1 \ \vdots \ e_n \ ] \qquad \ldots(1)$$

Since, $E(\varepsilon) = 0$, then the expression in (1) can be simply in the form of $b = Ax$

Where $b = [y_1 \ \vdots \ y_n \ ]$, $A = [1 \, x_{11} \ \ldots \ \ x_{1k} \ \vdots \vdots \ddots \quad \vdots \ 1 \, x_{n1} \ \ldots \ \ x_{nk} \ ]$, $x = [\beta_0 \ \vdots \ \beta_n \ ]$(the unknown solution)

The normal equation is simply expressed in the form;

$$X^T Y = (X^T X)\beta.$$

Where X^T defines transpose of matrix X

If we multiply through by inverse of the matrix $X^T X[i.e. (X^T X)^{-1}]$, we have;

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad \ldots(2)$$

where $\hat{\beta} = x = [\beta_0 \ \vdots \ \beta_n \ ]$

By evaluating the predicted value $(\hat{Y})$ for the mean of Y(since $Y = X\beta + \varepsilon$), where $E(\varepsilon) = 0$, we have;

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y.$$

$$\varepsilon = Y - \hat{Y}$$

Example: Consider the four pairs of (x, y) data points $\{(-1, -3), (0, -1), (1,1), (2,1)\}$

The system of the normal equation to find the best fit linear relationship for this data is of the form,

$b = Ax$ which can be expressed as;

$$[-3, -1, 1 \ , 1 \ ] = [[1, 1 \ ,1, 1 \ ], \ \ [-1, 0, 1, 2 \ \ ] \ ][\beta_{0,} \, \beta_1 \ ]$$

Applying the expression in (2);

$$X = [[1, 1 \ ,1, \ \ 1 \ ], \ \ [-1, 0, 1, 2 \ \ ] \ ], Y = [-3, -1, 1 \ , 1 \ ] \text{ and } \hat{\beta} = [\beta_{0,} \, \beta_1 \ ]$$

So,

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$= ([[1, -1, 1, 0] \ , [\ 1, 1, 1, 2] \ ][[1,1, 1, \qquad 1] \ , [-1, 0 \ ,1 \ ,2 \ ]])^{-1}[[1, -1,1, 0], [1,1,1, 2 \ ][-3, -1, 1, 1] \ ]$$

$$= [-1.2 \ , 1.4 \ ]$$

Therefore;

$$\beta_0 = -1.2, \ \ \beta_1 = 1.4$$

and the regression equation is of the form;

$$Y = -1.2 + 1.4X$$

Note:

- The coefficient of determination(R^2) for the linear regression is simply the square of the correlation between the vectors X and Y. We also could remember that correlation between X and Y expressed as $\rho_{(X,Y)} = \frac{Cov_{(X,Y)}}{\sigma_X \sigma_Y}$, where $Cov_{(X,Y)}$ is the covariance between X and Y and $\sigma_X \sigma_Y$ is the product of the standard deviation of X and Y respectively. Hence, the correlation between X and Y is obtained as 0.89442, and then $R^2 = 0.89442^2 = 0.8$
- Using R to solve the linear system, using the "summary" function, we obtained the result below;

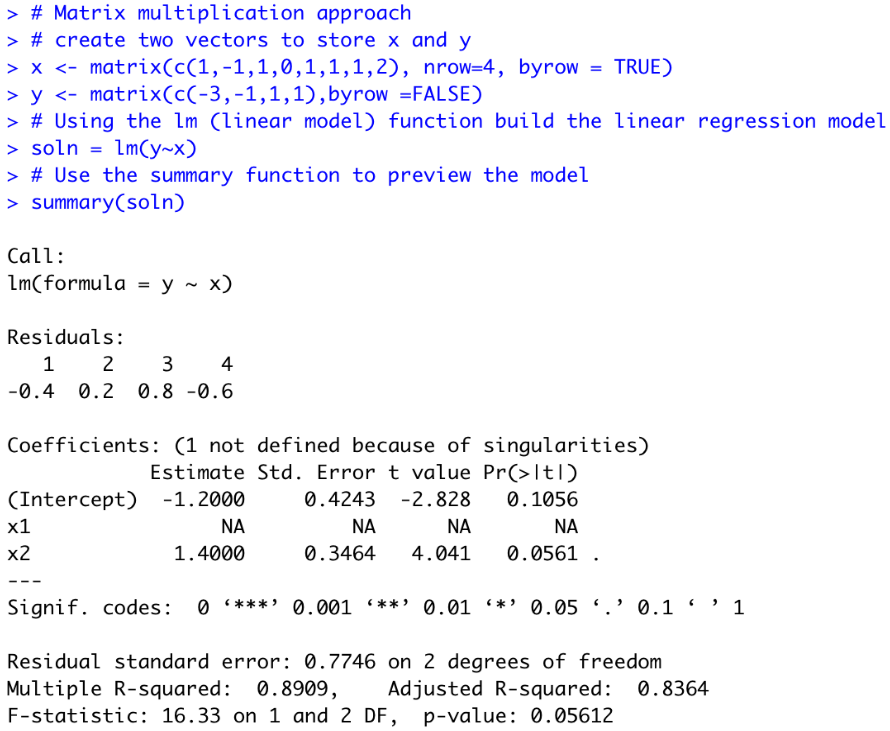
Implementation in R:

Having established sufficient theoretical background on linear regression, we now implement the above example in R to explain how the problem could be solved using a statistical programming package. The first approach uses the "lm" function in R, and the alternative method uses the matrix approach, as explained above.

- **Using the "lm" function in R**: we could simply create the X and Y vectors (data points) and apply the "lm" function to obtain the same result as the theoretical approach we obtained earlier(as shown below).

```
> # Create two vectors to store x and y
> x<-c(-1,0,1,2)
> y<-c(-3,-1,1,1)
> # Using the lm (linear model) function build the linear regression model
> soln = lm(y~x)
> # Use the summary function to preview the model
> summary(soln)

Call:
lm(formula = y ~ x)

Residuals:
   1    2    3    4 
-0.4  0.2  0.8 -0.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -1.2000     0.4243  -2.828   0.1056  
x             1.4000     0.3464   4.041   0.0561 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7746 on 2 degrees of freedom
Multiple R-squared:  0.8909,	Adjusted R-squared:  0.8364 
F-statistic: 16.33 on 1 and 2 DF,  p-value: 0.05612
```

- Alternatively, we could create a vector matrix by using the “**matrix()**” function in R, then apply the “lm” function to obtain the same result as shown below.



```
> # Matrix multiplication approach
> # create two vectors to store x and y
> x <- matrix(c(1,-1,1,0,1,1,1,2), nrow=4, byrow = TRUE)
> y <- matrix(c(-3,-1,1,1),byrow =FALSE)
> # Using the lm (linear model) function build the linear regression model
> soln = lm(y~x)
> # Use the summary function to preview the model
> summary(soln)

Call:
lm(formula = y ~ x)

Residuals:
   1    2    3    4 
-0.4  0.2  0.8 -0.6 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -1.2000     0.4243  -2.828   0.1056  
x1                NA         NA      NA       NA  
x2            1.4000     0.3464   4.041   0.0561 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7746 on 2 degrees of freedom
Multiple R-squared:  0.8909,	Adjusted R-squared:  0.8364 
F-statistic: 16.33 on 1 and 2 DF,  p-value: 0.05612
```

Working with Macroeconomic Variables using multiple linear regression

Case Study:

What is the association between GDP and other macroeconomic variables?

Data:

Data: Gross Domestic Product (U.S. Bureau of Economic Analysis)

Source: https://fred.stlouisfed.org/series/GDP

Note: The data for each macroeconomic variable considered was downloaded and merged into one single excel sheet to form one single table. Additional macroeconomic variables were added using data munging/wrangling techniques to determine how they contribute to our model.

Period Covered by Data: 04/01/1953 – 01/01/2021

Data Transformation & Processing:

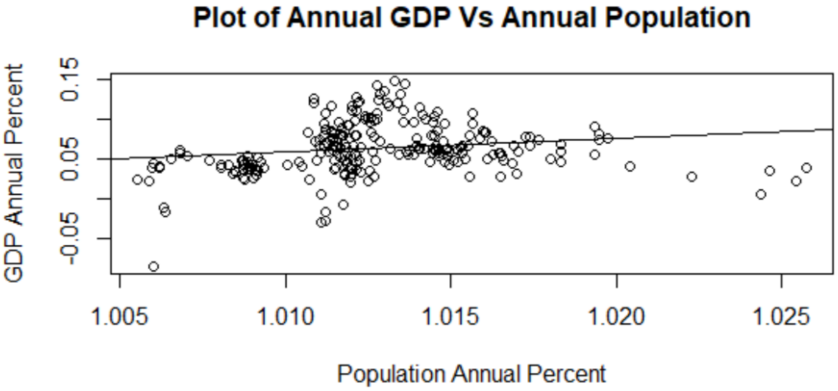
Missing values were handled using several methods as documented below;

- averaging (replacing the missing value(s) with the mean of the non-missing values -if no outlier)
- replacing the missing value(s) with the median of the non-missing values – if an outlier is present.
- in R, we can use the “mice package” to handle the missing value. This package is such that it uses some information from the other variables in the dataset to impute the missing data.

In this project, we used the “mice package”. A data frame is created for the missing values, and then the “mice package” imputed the missing data to ensure there was no longer any missing data (see R-code).

Exploratory Analysis:

The data consists of 272 records and 28 features. We are interested in modeling multiple independent variables to determine the relationships between these variables and the dependent variable (GDP). We also examined colinearity and interactions of the independent variables on our model. To aid in understanding the relationships between the variables we used scatterplots like the one shown below.



Based on our exploratory analysis and given that we are trying to predict a continuous variable, we concluded that our best choice for this part of the research is a linear regression model.

Base Model:

We start the regression analysis by initially modeling GDP with the variables listed below. Subsequently, after building the initial model, we added additional macroeconomic variables that could potentially improve the performance of the model.

Model:

Dependent Variable: GDP (Annual Percent) – GDP
Independent Variables: Population (Annual Percent) – Pop

Interest Rate(Annual Percent)- INTDSRUSM193
10-Year Treasury Constant Maturity Rate – DGS10
Disposable Personal Income(Annual Percent) – DSPI
Unemployment Rate(Annual Percent) – U2RATE
All Transactions House Price Index – USSTHPI

Model:

$$GDP = \beta_0 + \beta_1 Pop + \beta_2 INTDSRUSM193 + \beta_3 DGS10 + \beta_4 DSPI + \beta_5 U2RATE + \beta_6 USSTHPI$$

Where each variable represents the Annual percent returns and regression coefficients obtained are approximated to three decimal places.

Result:

Model:

$$GDP = -0.615 + 0.617 * Pop - 0.003 * INTDSRUSM193 + 0.015 * DGS10 + 0.716 * DSPI$$

$$-0.032 * U2RATE + 0.134 * USSTHPI$$

Table 1: Base model variable output

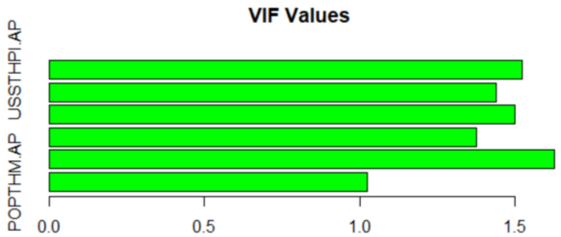
Input Variables	Estimate	VIF	R^2	Adjusted R^2	P-Value
Intercept	**-0.614682**				
Population	0.617380	1.023522			**
Interest Rate	-0.002543	1.628895			
Maturity Rate	0.015394	1.376404	0.8369	0.8332	**
Disposable Income	0.715986	1.499436			***
Unemployment Rate	-0.032274	1.441786			***
House Price Index	0.134140	1.525343			***

Key: *** - Highly significant: ** - Moderately Significant: * - Low Significant: No Significant (Empty)

- **p-value:** Disposable income, unemployment rate and House Price Index (HPI) are highly significant as shown in Table 1. However, both the population rate and the Maturity rate variable are moderately significant, whereas interest rate appears not to be significant.
- **Multicollinearity Check:** We obtained the Variance Inflation Factor (VIF) to check if multicollinearity exists among the independent variables. Research suggested that a VIF=1 implies no multicollinearity, VIF = 1.0 to 5.0 shows moderate collinearity, while VIF > 5 implies the

independent variables are highly correlated. In this analysis, we set the threshold for VIF to 2. Therefore, we assumed that VIF > 2 seems to be indicative of multicollinearity. In Table 1, we concluded that each independent variable is not correlated with the other since each VIF ≤ 2. Similarly, the VIF barplot below shows that none of the variables breached the threshold line (> 2). The Adjusted R^2 (83.7%) Implies that about 84% of the change in GDP is explained by the changes in the predictor variables.

Figure 1: VIF Bar-plot for base model variables



Improving the Model

In order to determine whether we can improve the performance of our model, we added the following additional variables one at a time to the base model:

- **Corporate Profit(CP):** the predictor variable is “Corporate Profit after Tax (without IVA and CCAdj) annual percent” (Corp).
- **10-Year Breakeven inflation rate (T10YIE):** the predictor variable is “10-Year Breakeven inflation rate annual percent” (T10YIE).

Process Followed to Add New Macroeconomic Variables to the Existing Dataset:

In order to merge the additional variable(s) into the existing data in R, we performed the following steps:

- Prepared a working data frame to correctly merge the new variable with the existing macroeconomic variables.
- Ensured that the “Date” column is set as the “primary key” so that when importing the new data, the “date” column is in date format rather than a factor format. This step was necessary because not all the variables are available for exactly the same timeframe.
- We inspected the new combined data to make sure everything that the data properly aligned. For example, by checking for missing values and filling them using the “mice package” if they exist.
- All the steps performed are documented in the accompanying R code available on the author’s GitHub page.

Extended Model 1- (Base Model Plus Corporate Profit Variable)

This model is an extension of the base model. This new model uses all the variables used for modeling the base model and an additional variable (Corporate Profit Variable).

The output of this new model is shown below:

Model:

$$GDP = -0.775 + 0.776 * Pop - 0.002 * INTDSRUSM193 + 0.015 * DGS10 + 0.702 * DSPI$$
$$+0.108 * U2RATE + 0.019 * USSTHPI + 0.019 * Corp$$

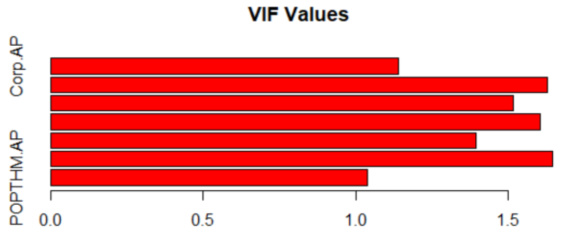
Table 2: Base model + additional variable output

Input Variables	Estimate	VIF	R^2	Adjusted R^2	P-Value
Intercept	**-0.774734**				
Population	0.775654	1.039206			**
Interest Rate	-0.002331	1.645682			
Maturity Rate	0.014681	1.394145	0.8436	0.8394	**
Disposable Income	0.702367	1.605807			***
Unemployment Rate	0.108125	1.515550			***
House Price Index	0.018849	1.628861			***
Corporate Profit	0.018849	1.139130			***

Key: *** - Highly significant: ** - Moderately Significant: * - Low Significant: No Significant (Empty)

- **p-value:** Disposable income, unemployment rate, House price index, and corporate profit (newly added) are highly significant as shown in Table 2. However, both the population rate and maturity rate variable are moderately significant, whereas interest rate appears not to be significant.
- **Multicollinearity Check:** From Table 2, each VIF<2 shows no multicollinearity based on the threshold. Similarly, the VIF barplot in figure 2 shows that none of the variables cut across the threshold line(> 2). We observed that the R^2 (84.4%) and the adjusted R^2 (83.9%) increases slightly than what we had in Table 1. This shows that the new variable slightly improves the model fit.

Figure 2: VIF Bar-plot for base model + additional variable



Extended Model 2- (Extended Model 1 Plus 10-Year Breakeven Inflation Rate)

This model is an extension of the extended model 1 above. This new model uses all the variables used for modeling the extended model 1 above and an additional variable (10-Year Breakeven Inflation Rate).

The output of this new model is shown below:

Model:

$$GDP.AP = -1.604 + 1.612 * Pop + 0.008 * INTDSRUSM193 + 0.002 * DGS10 + 0.263 * DSPI$$
$$-0.024 * U2RATE + 0.120 * USSTHPI + 0.009 * Corp - 0.001 * T10YIE$$

Table 3: Base model + additional variables output

Input Variables	Estimate	VIF	R^2	Adjusted R^2	P-Value
Intercept	**-1.6038198**				
Population	1.6116474	1.325264			**
Interest Rate	0.0075053	2.325218			*
Maturity Rate	0.0020165	1.843710	0.9054	0.8935	
Disposable Income	0.2629742	1.414084			***
Unemployment Rate	-0.0237378	1.916732			***
House Price Index	0.1199530	2.231138			***
Corporate Profit	0.0093735	4.263010			
Inflation Rate	-0.0006294	4.446946			

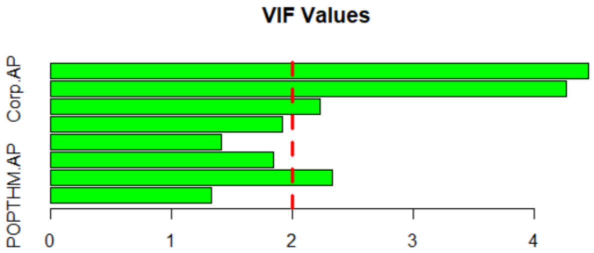
Key: *** - Highly significant: ** - Moderately Significant: * - Low Significant: No Significant (Empty)

- **p-value:** We see that the precision of our estimated coefficients reduces. Disposable income, unemployment rate, and House price index remain highly significant, while population rate is moderately significant and interest rate has a low significant level, as shown in Table 3. However, maturity rate, corporate profit, and Inflation rate appear not to be statistically significant. This could be traced to the multicollinearity effect when new variables are added.

  NOTE: Since multicollinearity affects p-values, and regression coefficients, we cannot trust the p-values obtained even though it does not influence our predictions.

- **Multicollinearity Check:** From Table 3, based on the VIF threshold that we set for this research, only population, maturity rate, disposable income, and unemployment rate variables have VIF$<$ 2 which indicates no multicollinearity. Interest rate and House price index with VIF$\sim \leq 2$ shows very little or no multicollinearity. However, Corporate profit and Inflation rate variables show some level of multicollinearity with VIF> 2. Although, research suggested that a VIF≤ 5 has moderate collinearity for research purposes. We observed that the R^2 (90.5%) and the adjusted R^2 (89.4%) increases compared to what we have in Table 2. This shows that the new variable improves the model fit even though with some level of moderately acceptable multicollinearity.

Figure 3: VIF Bar-plot for base model + additional variables



Conclusion:

In this essay, we examined how GDP could be predicted using other macroeconomic variables. To achieve this objective, we used multiple linear regression analysis. We first built a based model and then extended the base model to inculcate additional variables to determine whether we could improve the performance of our base model. We also evaluated how each additional variable impacted our regression fit by comparing the VIF's and the p-values at a 5% significant level. The multiple linear regression model appears to be a suitable model for determining a linear relationship between dependent and independent features.

References:

- FRED Economic Data : Data Source
- Felipe Rego : Interpreting Linear Regression Model in R
- Author's Github: GitHub