VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY
INTERNATIONAL UNIVERSITY
DEPARTMENT OF MATHEMATICS

STATISTICS, PROJECT REPORT
THE IMPACTS OF MACROECONOMICS INDICATORS ON U.S. GROSS DOMESTIC PRODUCT

Lecturer: Dr. Nguyen Minh Quan

Group Members: Ton Nu Trieu Man - MAMAIU20037
Tran Viet Hang - MAMAIU18079
Le Trong Tan - MAMAIU20032

Ho Chi Minh City, Vietnam
December 2022

Contents

List of Figures

Chapter 1

Introduction

The COVID-19 pandemic exerts a catastrophic impact on the global economy. A major concern in a crisis is the influence on business which consequently determines the unemployment, debt level, financial markets, etc in a country. It is difficult to identify economies that have evaded the detrimental effects of COVID-19. However, some nations have seen a smaller decline in economic growth than others. Therefore, knowing the prospective economic indicators can help to afford economic remedies.

Gross Domestic Product (GDP) has been used as one of the valid measures to assess a country's level of wealth. By its definition, "GDP measures the monetary value of final goods and services - that is, those that are bought by the final user - produced in a country in a given period of time (say a quarter or a year)" *(Callen, 2019)*. The GDP growth rate is frequently seen as a sign of the economy's overall health. In general, growth in GDP is seen as a positive indicator of the health of the economy. It is influenced by various macroeconomic indicators such as consumption, investment, government spending, exports, imports, etc. The decision-maker and the analyst's subjective judgment will have a role in how these parameters are chosen.

For the above reasons, it is very important to understand the GDP as well as which indicators determine it. This paper aims to investigate the determinants of quarterly GDP in the US by mainly applying Multi-linear Regression with the Ordinary Least Square (OLS) method. Data for this research project is extracted from Federal Reserve Economic Data, which are macroeconomic indicators recorded from the first quarter of 1980 to the third quarter of 2022. The implementation of the methodology on the dataset would previously consider five assumptions of the OLS method so that the dataset can be treated as trustworthy.

Through that, it also measures the level of influence of these determinants as well as estimates the implicit GDP in the next several quarters. Finally, the result and findings from this research could introduce some implications and suggestions for further studies. It could be employed as

a reference for researchers and investors to understand the future prospects of the US economy, thus, make investment decisions. In other words, the findings in this study could assist their decision-making for investment strategies or development plans. The research aims to answer the following research questions:

1. What are the determinants of the GDP in the US economy?
2. How do the attributes contribute to the GDP in the US and what is the importance of the attributes?

Chapter 2

Methodology

2.1 Stepwise Regression

This research aims to figure out explanatory variables which contribute significantly. Stepwise regression is a method of fitting regression models, in which the feature selection process is conducted as follows. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion. Usually, this takes the form of a forward, backward, or combined sequence of F-tests or t-tests.

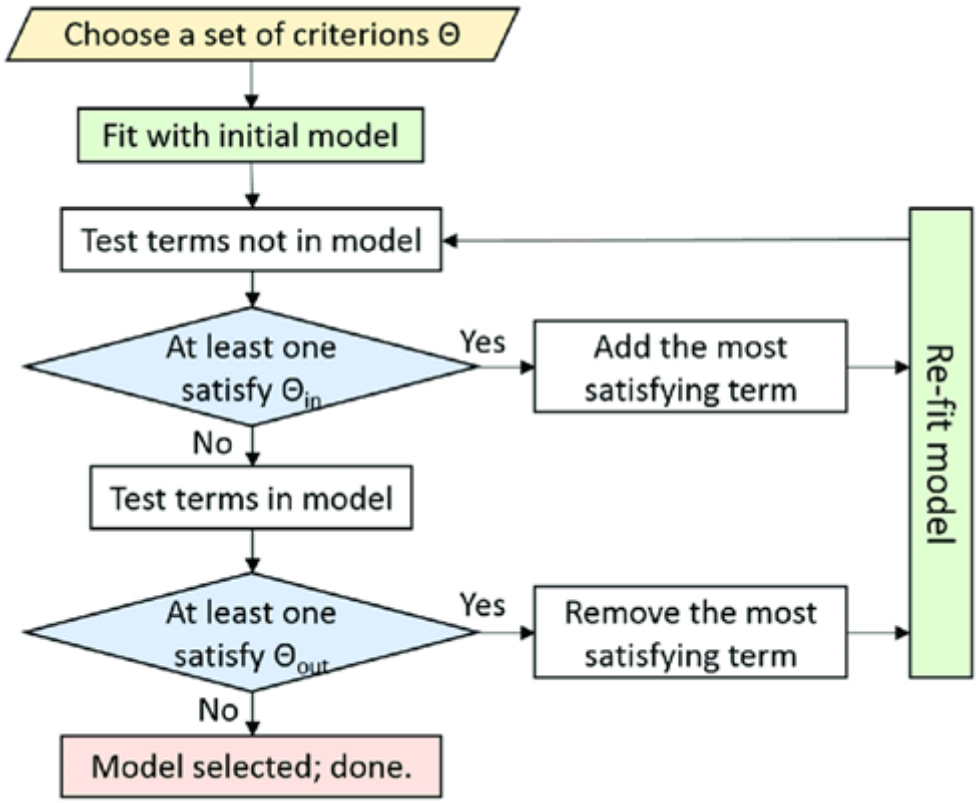

Figure 2.1: A schematic diagram of Stepwise Regression

2.2 Ordinary Least Squares method

2.2.1 Ordinary Least Squares method for Simple Linear Regression

The Simple Linear Regression problem is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where

- the index i runs over the observations $i = 1, .., n$
- Y_i is the dependent variable, the regressand, or simply the left-hand variable
- X_i is the independent variable, the regressor, or simply the right-hand variable
- $Y = \beta_0 + \beta_1 X$ is the population regression line also called the population regression function
- β_0 is the intercept of the population regression line.
- β_1 is the slope of the population regression line.
- ϵ_i is the error term.

The solution of the proposed problem is the solution of the following optimization problem: minimize:

$$\text{minimize: } SS = \sum_{i=1}^{n} (\epsilon_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

where SS is called the sum of squared residuals.

The derivatives of SS with respect to β_0 and β_1 are taken, then the solutions would be obtained as the following:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{x}$$

2.2.2 Ordinary Least Squares method for Multiple Linear Regression

This paper aims to investigate the impacts of different macroeconomics factors on the GDP, thus the Multiple Linear Regression Model would be introduced.

In particular, the Multiple Linear Regression problem is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \epsilon_i, \forall i = \overline{1, n}.$$

The designations are similar to those in the simple regression model:

- Y_i is the i^{th} observation in the dependent variable. Observations on the k regressors are denoted by $X_{1i}, X_{2i}, ..., X_{ki}\text{and}\epsilon_i$ is the error term.
- The average relationship between Y and the regressors is given by the population regression line

$$\mathbb{E}(Y_i|X_{1i} = x_1, X_{2i} = x_2, X_{3i} = x_3, ..., X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k$$

- β_0 is the intercept; it is the expected value of Y when all $X's$ equal 0. $\beta_j, j = 1, ..., k$ are the coefficients on $X_j, j = 1, ..., k$. β_1 measures the expected change in Y_i that results from a one unit change in X_{1i} while holding all other regressors constants.

The Multiple Linear Regression problem can be written in the matrix form:

$$y = X\beta + \epsilon$$

where y and ϵ are $n \times 1$ vectors of the response variables and the errors of the n observations, and X is an $n \times p$ matrix of regressors, also sometimes called the design matrix, whose row i is x_i^T and contains the i^{th} observations on all the explanatory variables.
Similarly to the Simple Linear Regression problem, the solution of the generalized problem would be implied by solving:

$$\text{minimize } SS = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = (y - X\beta)^T (y - X\beta)$$

That is:

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

The goodness-of-fit of the OLS regression is commonly assessed by the coefficient of determination R^2:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{ESS}{TSS}$$

where:

- TSS is the total sum of squares for the dependent variable, TSS $= \sum (y_i - \bar{y})^2$

- ESS is the estimated sum of squares for the dependent variable, ESS $= \sum(\hat{y}_i - \bar{y})^2$
- TSS = RSS + ESS, RSS = $\sum_i \hat{\epsilon_i}^2$

In that case, R^2 will always be a number between 0 and 1, with values closer to 1 indicating a better degree of fit.

2.2.3 Ordinary Least Squares assumptions

There are several different frameworks in which the linear regression model can be cast in order to make the OLS technique applicable. Each of these settings produces the same formulas and the same results. The only difference is the interpretation and the assumptions which have to be imposed in order for the method to give meaningful results. This research aims to verify five assumptions of Ordinary Least Squares (OLS) as follows:

Assumption of Strict exogeneity

The errors in the regression should have conditional mean zero:

$$E(\epsilon_i) = 0$$

The exogeneity assumption is critical for the OLS theory. If it holds then the regressor variables are called exogenous. If a constant term is included in the regression equation, this assumption will never be violated. First, R^2, defined as ESS/TSS can be negative, implying that the sample average, 'explains' more of the variation in y than the explanatory variables. Second, and more fundamentally, a regression with no intercept parameter could lead to potentially severe biases in the slope coefficient estimates.

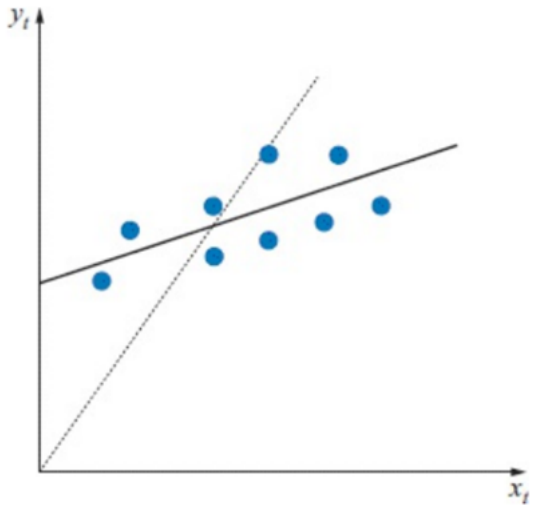

Figure 2.2: Effect of no intercept on a regression line

Assumption of Homoscedasticity

The error term has the same variance σ^2 in each observation. When this requirement is violated this is called heteroscedasticity, in such case a more efficient estimator would be weighted least squares. If the errors have infinite variance then the OLS estimates will also have infinite variance (although by the law of large numbers they will nonetheless tend toward the true values so long as the errors have zero mean). In this case, robust estimation techniques are recommended.

$$Var(\epsilon_i) = \sigma^2 < \infty$$

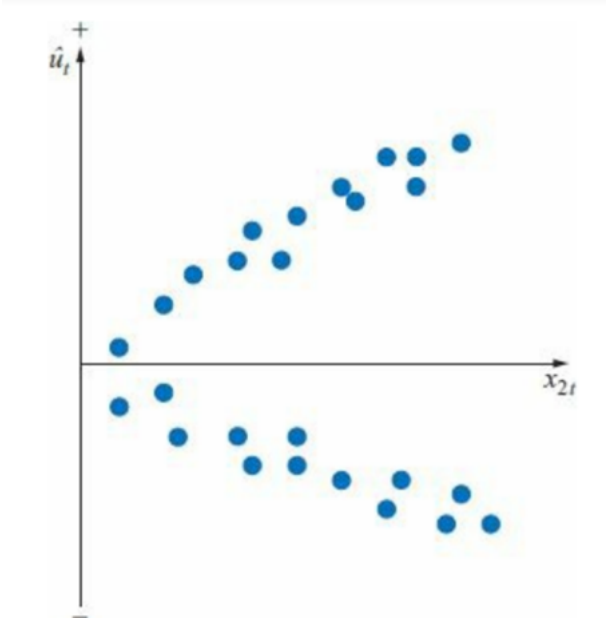

Figure 2.3: Graphical illustration of heteroscedasticity

This assumption will be tested by using the White's test in this paper.

Assumption of Non-autocorrelation

The errors are uncorrelated between observations: $\mathbb{E}(\epsilon_i \epsilon_j | X) = 0$, or $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$ This assumption may be violated in the context of time series data, panel data, cluster samples, hierarchical data, repeated measures data, longitudinal data, and other data with dependencies. In such cases generalized least squares provides a better alternative than the OLS.

This assumption will be tested by using the Durbin Watson test.

Assumption of non-stochastic independent variables

$$cov(x_i, \epsilon_i) = 0$$

or x_i are non-stochastic. Fortunately, the OLS is consistent and unbiased even in the case of stochastic x_i.

Assumption of Normally Distributed Residuals

The errors have normal distribution conditional on the regressors:

$$\epsilon|X \sim \mathcal{N}(0, \sigma^2 I_n)$$

This assumption is not needed for the validity of the OLS method, although certain additional finite-sample properties can be established in case when it does (especially in the area of hypothesis testing). This condition can be tested by the Jarque-Bera test.

In addition, if the assumptions 1-4 hold, then the estimators determined by OLS will have a number of desirable properties, and are known as best linear unbiased estimators (BLUE). That is:

- 'Estimator' – $\hat{\beta}$ are estimators of the true value of β.
- 'Linear' – $\hat{\beta}$ with β_i are linear estimators – that means that the formulae for β_i are linear combinations of the random variables (in this case, y).
- 'Unbiased' – on average, the actual values of $\hat{\beta}$ and will be equal to their true values.
- 'Best' – means that the OLS estimator has minimum variance among the class of linear unbiased estimators; the Gauss–Markov theorem proves that the OLS estimator is best by examining an arbitrary alternative linear unbiased estimator and showing in all cases that it must have a variance no smaller than the OLS estimator.

An implicit assumption of Autocorrelation that the explanatory variables are not correlated with one another would be under the consideration in this paper. The assumption verification can employ variance inflation factors (VIF). However, in this project, the authors have excluded pairs of highly correlated features, and VIF would not be calculated in this paper.

2.3 Hypothesis Testing

2.3.1 t-test

This project aims to test whether an independent variable has significant impact on the target, or the dependent variable, thus the paper would consider the following hypothesis test for coefficients of each independent variable: $H_0 : \beta_i = 0$ vs $H_a : \beta_i \neq 0$ for each $i = 0, 1, 2, ...$ t-test would be applied in this paper as it is primarily used with samples with limited information, and, thus, the

volatility of the whole population is unknown.

Consider the sample where n is the sample size, $\bar{x}$ and $\bar{Y}$ are the sample means, S is the sample standard deviation, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{xY} = \sum_{i=1}^{n}(x_i - \bar{X})(Y_i - \bar{Y})$. Denote $B = \frac{S_{xY}}{S_{xx}}$

$$\sqrt{\frac{(n-2)S_{xx}}{SS}} \cdot (B - \beta_i) \sim T_{n-2}$$

The test statistic in this case:

$$\text{TS: } = \sqrt{\frac{(n-2)S_{xx}}{SS}} \cdot |B|$$

The hypothesis test would be rejected for any test statistics TS $> t_{\alpha/2,n-2}$, or for any p-value $2 \cdot \mathbb{P}(T_{n-2} > TS) < \alpha$, where α is the significance level.

This process would omit factors with insignificant variables (with the assumption that its coefficient equals zero) by eliminating those with high p-value (larger than the level of significance). The remaining factors from this process would be estimated confidence intervals for each.

For α significance level, the confidence interval of the coefficient of individual factor would be obtained as following:

$$\left(B - \sqrt{\frac{SS}{(n-2)S_{xx}}}t_{\alpha/2,n-2}, B + \sqrt{\frac{SS}{(n-2)S_{xx}}}t_{\alpha/2,n-2}\right)$$

2.3.2 White's test

Heteroskedasticity and Homoscedasticity

The error term of our regression model is homoskedastic if the variance of the conditional distribution of ϵ_i given X_i, $Var(\epsilon_i|X_i = x)$ is a constant for all observations in the sample

$$Var(\epsilon_i|X_i = x) = \sigma^2, \forall i = 1, 2, ..., n$$

If instead there is dependence of the conditional variance of ϵ_i on X_i , the error term is said to be heteroskedastic.

$$Var(\epsilon_i|X_i = x) = \sigma_i^2, \forall i = 1, 2, ..., n$$

White's test

White's test is a statistical test for homoscedasticity. The hypothesis test would be as following: $H_0 : \sigma_i^2 = \sigma^2, \forall i = 1, ..., n$ vs $H_a : \exists i \in [1, n], \sigma_i^2 \neq \sigma^2$

The process conducting White's test:

- Assume that the regression model estimated is of the standard linear form, e.g.

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \epsilon_t$$

- Then run the auxiliary regression

$$\hat{\epsilon_t}^2 = \alpha_1 + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \alpha_4 x_{2t}^2 + \alpha_5 x_{3t}^2 + \alpha_6 x_{2t} x_{3t} + v_t$$

  where v_t is a normally distributed disturbance term independent of ϵ_t.

This approach is known as a Lagrange Multiplier (LM) test, which centers around the value of R^2 for the auxiliary regression. If one or more coefficients in an auxiliary model is statistically significant, the value of R^2 for that equation will be relatively high, while if none of the variables is significant, R^2 will be relatively low. The LM test would thus operate by obtaining R^2 from the auxiliary regression and multiplying it by the number of observations, T.

$$TR^2 \sim \chi^2(m)$$

where m is the number of regressors in the auxiliary regression (excluding the constant term).

For the LM test, if the χ^2-test statistic is greater than the corresponding value from the statistical table then reject the null hypothesis that the errors are homoscedastic. In other words, the null hypothesis would be rejected for any significance level larger than the p-value.

2.3.3 Durbin-Watson test

The Durbin-Watson test aims to detect the autocorrelation between error terms. Consider a regression of the time t error on its previous value:

$$u_t = \rho u_{t-1} + v_t$$

where $v_t \sim \mathcal{N}(0, \sigma_v^2)$

The test would test the hypothesis: $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$

The test statistic would be obtained as:

$$DW = \frac{\sum_{t=2}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=2}^{T}\hat{u_t}^2}$$

and $DW \approx 2(1-\hat{\rho})$, where $\hat{p}$ is the estimated correlation coefficient that would have been obtained from $u_t = \rho u_{t-1} + v_t$, $-1 \leq \hat{p} \leq 1$. Thus, test statistic DW would take values from 0 to 4, $0 \leq DW \leq 4$.

Consider:

- $\hat{p} = 0$, then $DW = 2$. This is the case where there is no autocorrelation in the residuals. In other words, the null hypothesis would not be rejected if DW is near 2.
- $\hat{p} = 1$, then $DW = 0$.This corresponds to the case where there is perfect positive autocorrelation in the residuals.
- $\hat{p} = -1$, then $DW = 4$.This corresponds to the case where there is perfect negative autocorrelation in the residuals.

2.3.4 Jarque Bera test

The Jarque Bera test is applied in this problem to test the normality of the error term, by examining its third and fourth moments - that is, skewness and kurtosis. A normal distribution is not skewed and is defined to have a coefficient of kurtosis of 3.
The test statistic:

$$W = T\left[\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right]$$

where $b_1 = \frac{\mathbb{E}(u^3)}{(\sigma^2)^{3/2}}$ and $b_2 = \frac{\mathbb{E}(u^4)}{(\sigma^2)^2}$, denoting the errors by u and their variance by σ^2.

The null hypothesis is of normality, and this would be rejected if the residuals from the model were either significantly skewed or leptokurtic/platykurtic (or both).

Chapter 3

Data Analysis

The data is collected quarterly from the Federal Reserve Bank's website with time ranging from Q1 1980 to Q3 2022 ($n = 171$). There is 1 dependent variable (GDP) and 19 independent variables whose description will be shown in detail below.

- *GDP*: Gross Domestic Product (Billions of Dollars)
- *MMMFFAQ027S*: Money Market Funds; Total Financial Assets, Level (Millions of Dollars)
- *ASTMA*: All Sectors; Total Mortgages; Asset, Level (Millions of Dollars)
- *BOGZ1FL072052006Q*: Interest Rates and Price Indexes; Effective Federal Funds Rate (Percent), Level
- *ROWFDIQ027S*: Rest of the World; Foreign Direct Investment in U.S.; Asset (Current Cost), Transactions (Millions of Dollars)
- *BOGZ1FL594090005Q*: Pension Funds; Total Financial Assets, Level (Millions of Dollars)
- *FGCCSAQ027S*: Federal Government; Consumer Credit, Student Loans; Asset, Level (Millions of Dollars)
- *TLAACBQ158SBOG*: Total Assets, All Commercial Banks (Millions of Dollars)
- *BOGZ1FA895050005Q*: All Sectors; Total Capital Expenditures, Transactions (Millions of Dollars)
- *FGLBAFQ027S*: Federal Government; Net Lending (+) or Borrowing (-) (Financial Account), Transactions (Millions of Dollars)
- *LES1252881600Q*: Employed full time: Median usual weekly real earnings: Wage and salary workers: 16 years and over (CPI Adjusted Dollars)

- *LREM64TTUSQ156S*: Employment Rate: Aged 15-64: All Persons for the United States (Percent)

- *RSAHORUSQ156S*: Homeownership Rate in the United States (Percent)

- *NROU*: Noncyclical Rate of Unemployment (Percent)

- *A067RL1Q156SBEA*: Real Disposable Personal Income (Percentage change from the preceding period)

- *GFDEBTN*: Federal Debt: Total Public Debt (Millions of Dollars)

- *GGSAVE*: Gross Government Saving (Billions of Dollars)

- *NETEXP*: Net Exports of Goods and Services (Billions of Dollars)

- *CCUSMA02EZQ618N*: Currency Conversions: US$ Exchange Rate: Average of Daily Rates: National Currency:USD for the Euro Area (19 Countries) (Euros)

- *M2V*: Velocity of M2 Money Stock (Ratio)

	GDP	MMMFFAQ027S	ASTMA	BOGZ1FL072052006Q	ROWFDIQ027S	BOGZ1FL594090005Q	FGCCSAQ027S	TLAACBQ158SBOG	BOGZ1FA895050005Q
count	171.000000	1.710000e+02	1.710000e+02	171.000000	171.000000	1.710000e+02	1.710000e+02	171.000000	1.710000e+02
mean	11377.838667	1.839531e+06	8.569869e+06	4.389474	155555.397661	1.114659e+07	3.136796e+05	6.080117	3.275149e+06
std	6102.187837	1.433792e+06	5.424841e+06	4.058849	156797.198801	7.568698e+06	4.720220e+05	5.379117	1.683796e+06
min	2789.842000	6.129200e+04	1.357614e+06	0.070000	-293968.000000	1.560558e+06	0.000000e+00	-11.400000	8.388860e+05
25%	6009.924500	4.940830e+05	3.758682e+06	0.885000	34436.000000	4.308769e+06	0.000000e+00	3.000000	1.724613e+06
50%	10598.020000	1.865022e+06	7.057091e+06	3.980000	110320.000000	9.649967e+06	6.668300e+04	6.700000	3.263405e+06
75%	15955.545500	2.969142e+06	1.384548e+07	6.525000	261354.000000	1.652744e+07	5.130760e+05	8.750000	4.344534e+06
max	25698.960000	5.205455e+06	1.905848e+07	19.100000	956604.000000	2.766260e+07	1.484310e+06	43.700000	7.534022e+06

Figure 3.1: Sample summary of GDP and 8 independent variables

Since some of the variables in the data set have different units, it is better to standardize data before making any visualization so that we could recognize the pattern as well as the trend of GDP and all macroeconomic factors more clearly and accurately.

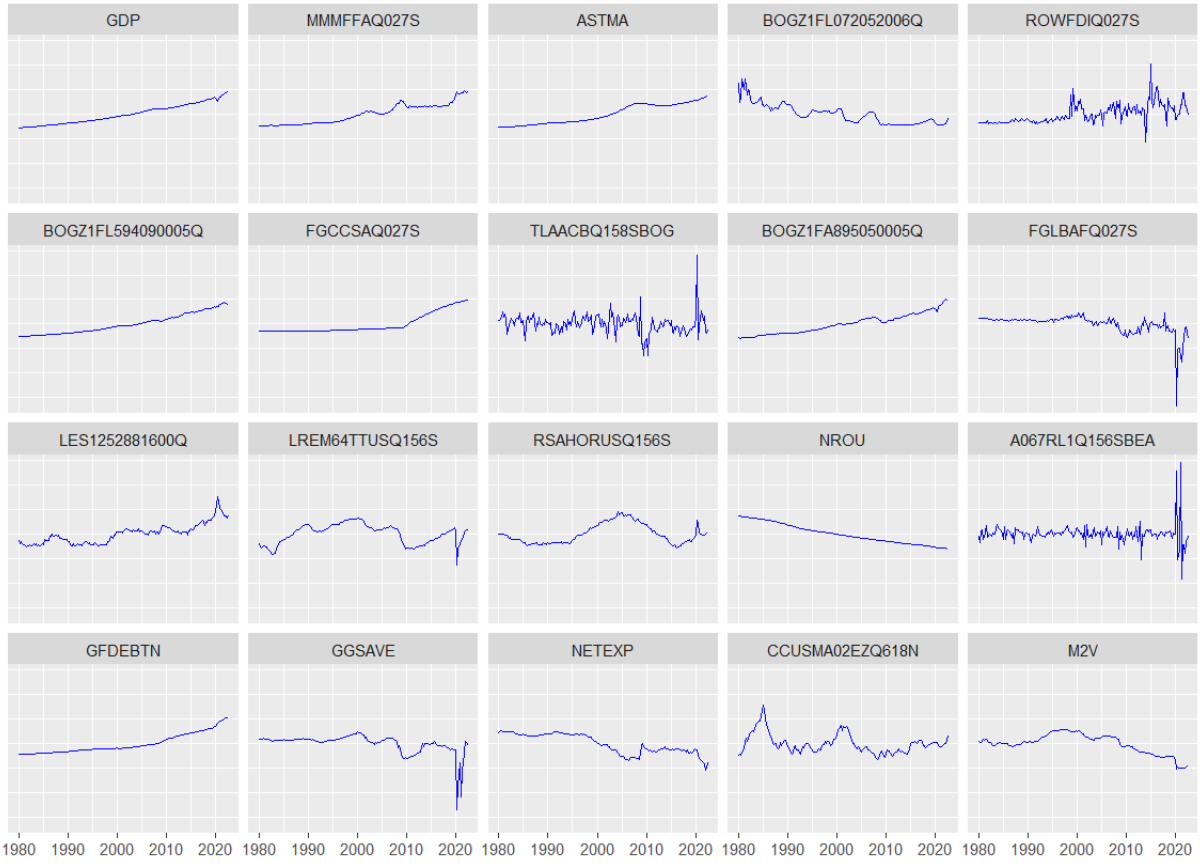

Figure 3.2: U.S GDP & Macroeconomic Factors Trend from 1980 to 2022

Overall, the GDP of the US witnessed a slight increase from Q1 1980 (2.7 trillion dollars) to Q3 2022 (25.6 trillion dollars). The majority of macroeconomic factors also experienced an upward trend as GDP went up over time. While MMMFFAQ027S, ASTMA, BOGZ1FL594090005Q, FGCCSAQ027S, BOGZ1FA895050005Q, GFDEBTN increased slowly, LES1252881600Q, LREM64TTUSQ156S, RSAHORUSQ156S, CCUSMA02EZQ618N increased but showed small fluctuations over time. In contrast, BOGZ1FL072052006Q, NROU, GGSAVE, NETEXP, M2V decreased slightly between 1980 and 2022. It is interesting that there were some volatile variables, which are TLAACBQ158SBOG, ROWFDIQ027S, FGLBAFQ027S and A067RL1Q156SBEA.

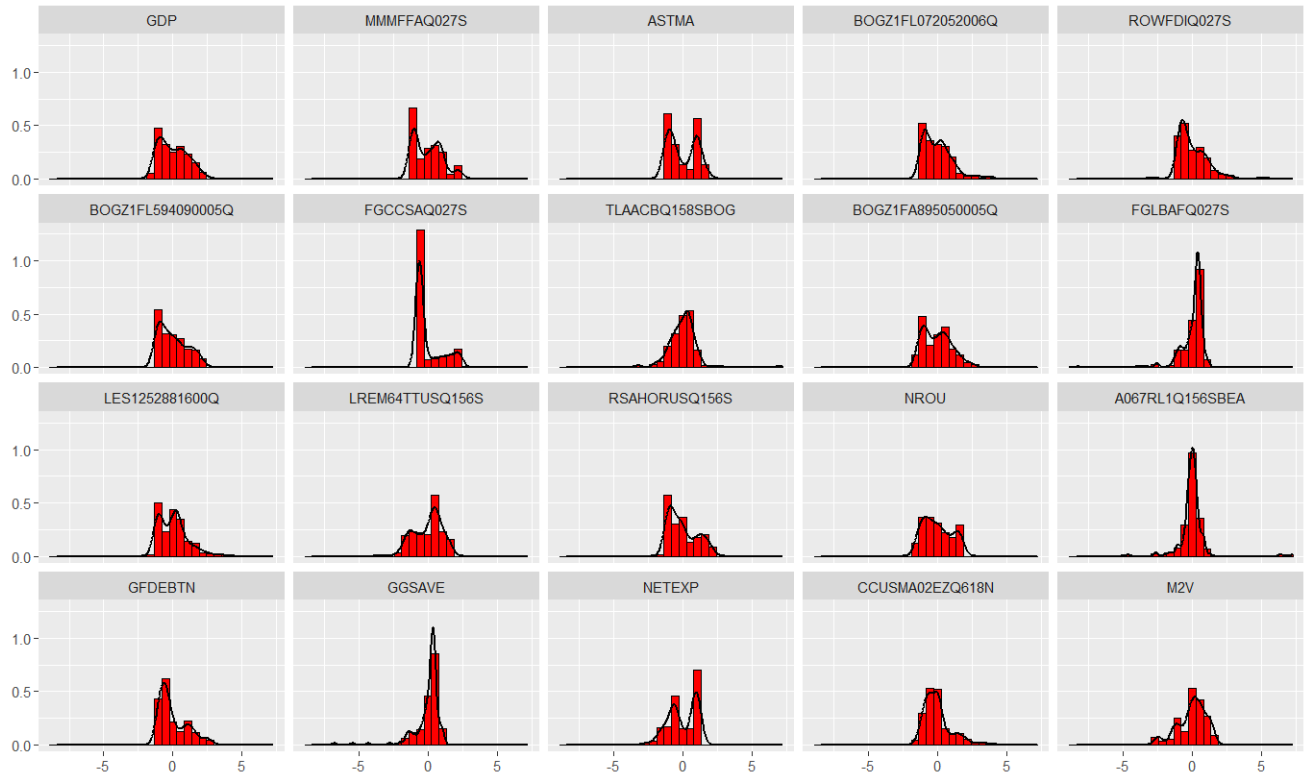

Figure 3.3: U.S GDP & Macroeconomic Factors Histogram from 1980 to 2022

Chapter 4

Results and Discussion

4.1 Results

Feature selection process is conducted and obtains the results as follows.

	CCUSMA02EZQ618N	LREM64TTUSQ156S	RSAHORUSQ156S	NROU
count	171.000000	171.000000	171.000000	171.000000
mean	-0.139581	4.251338	4.183116	1.652078
std	0.154526	0.034891	0.025563	0.101598
min	-0.446004	4.133858	4.144721	1.489704
25%	-0.250561	4.223684	4.160444	1.565812
50%	-0.162629	4.261835	4.178992	1.643378
75%	-0.079219	4.275493	4.204692	1.737720
max	0.380580	4.307664	4.239887	1.827347

Figure 4.1: Description of selected features using featurewiz package.

	CCUSMA02EZQ618N	LREM64TTUSQ156S	RSAHORUSQ156S	NROU
CCUSMA02EZQ618N	1.000000	0.044459	0.172945	0.355500
LREM64TTUSQ156S	0.044459	1.000000	0.164501	0.017672
RSAHORUSQ156S	0.172945	0.164501	1.000000	0.296899
NROU	0.355500	0.017672	0.296899	1.000000

Figure 4.2: The correlation matrix of selected features.

This project considers the log-scale of the historical data and aims to run the regression with independent variables, thus highly correlated features would be eliminated from the data set (using the *featurewiz* package in Python). Features with correlation higher than 0.7 would be excluded

from the data set.

From the obtained results, macroeconomics indicators which can contribute to the OLS Regression with p-value less than 1% remained after the Stepwise Regression. The p-values of Noncyclical Rate of Unemployment, Employment Rate: Aged 15-64: All Persons for the United States, and Currency Conversions: US$ Exchange Rate: Average of Daily Rates: National Currency:usd for the Euro Area (19 Countries) are 3.75684×10^{-134}, 5.60041×10^{-12}, and 6.80781×10^{-5}, respectively. In other words, for 1% significance level, the hypothesis testing on each features is rejected, or these features can contribute to the regression with 1% level of significance.

```
Add  NROU                              with p-value 3.75684e-134
============================================================
Add  LREM64TTUSQ156S                   with p-value 5.60041e-12
============================================================
Add  CCUSMA02EZQ618N                   with p-value 6.80781e-05
============================================================
```

Figure 4.3: Result of the Stepwise Regression (features with estimated p-value smaller than 1%)

The final result of OLS Regression run on Python is shown below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    GDP   R-squared:                       0.993
Model:                            OLS   Adj. R-squared:                  0.993
Method:                 Least Squares   F-statistic:                     6647.
Date:                Thu, 22 Dec 2022   Prob (F-statistic):          8.87e-144
Time:                        09:12:11   Log-Likelihood:                236.51
No. Observations:                 136   AIC:                            -465.0
Df Residuals:                     132   BIC:                            -453.4
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           16.1056      0.469     34.352      0.000      15.178      17.033
NROU            -6.1759      0.051   -121.541      0.000      -6.276      -6.075
CCUSMA02EZQ618N  0.1037      0.025      4.114      0.000       0.054       0.154
LREM64TTUSQ156S  0.7762      0.104      7.445      0.000       0.570       0.982
==============================================================================
Omnibus:                        0.111   Durbin-Watson:                   0.040
Prob(Omnibus):                  0.946   Jarque-Bera (JB):                0.274
Skew:                          -0.010   Prob(JB):                        0.872
Kurtosis:                       2.781   Cond. No.                         608.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 4.4: OLS Regression result, considering results of the Stepwise Regression

Using 70% of the collected data for the train data set, the model demonstrated the coefficient of determinant $R^2 = 0.993$. The movement of GDP in the United States in this regression would follow an equation:

$$\begin{aligned}\log(GDP) =& 16.1056 - 6.1759\log(NROU) + 0.1037\log(CCUSMA02EZQ618N) \\ &+ 0.7762\log(LREM64TTUSQ156S)\end{aligned}$$

The 95% confidence interval of coefficient of each explanatory variable:

- The intercept: $\beta_0 \in (15.178, 17.033)$
- The Noncyclical Rate of Unemployment $\beta_1 \in (-6.276, -6.075)$
- The Employment Rate: Aged 15-64: All Persons for the United States: $\beta_2 \in (0.054, 0.154)$
- The Currency Conversions: US$ Exchange Rate: Average of Daily Rates: National Currency:usd for the Euro Area (19 Countries): $\beta_3 \in (0.570, 0.982)$

The implied mean squared error (MSE) and mean absolute error (MAE) on train and test data set:

- MSE on train data set: 0.00180720861617508
- MSE on test data set: 0.005501422142687502
- MAE on train data set: 0.03557742369115147
- MAE on test data set: 0.06910106639643797

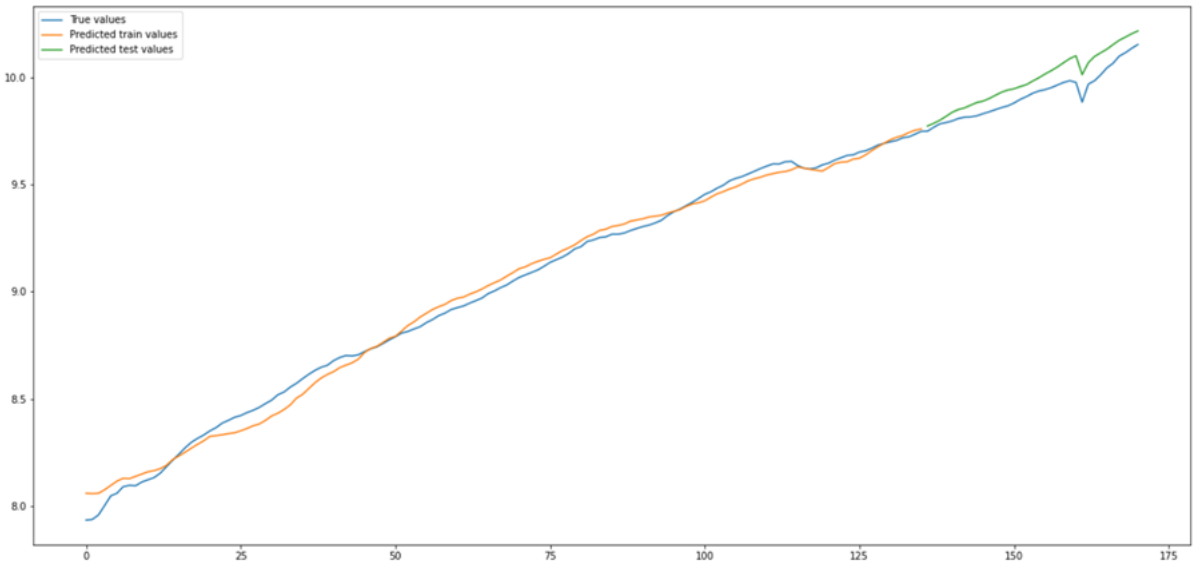

Figure 4.5: The graphical illustration of the results

However, for the $\alpha = 1\%$ level of confidence ($t_{0.005,169} = 2.61$), the OLS assumptions are not all satisfied on this data set, and assumptions 1-4 do not all hold, thus the estimator $\hat{\beta}$ is not considered as BLUE. In particular:

1. The assumption of Strict exogeneity is satisfied because the model has added a constant.

2. The assumption of Homoscedasticity is not satisfied by using White's test. According to the White's test run on Python, the test statistic is 97.88235474336877, implying the test statistic p-value of $4.2155547262464223 \times 10^{-17}$. Thus, for any significance level higher than the implied p-value, the null hypothesis of Homoscedasticity is rejected.

3. The assumption of Non-autocorrelation is not satisfied due to the low level of Durbin-Watson test statistic, in this case, of 0.040.

4. The assumption of non-stochastic independent variables is satisfied as mentioned.

5. The assumption of Normally Distributed Residuals is satisfied by using the Jarque-Bera test - that is, the Jarque Bera is 0.274 and Prob(JB) is 0.872.

Regarding the multicollinearity between variables, as mentioned, all pairs of features have correlation lower than 0.7, thus, the problem would not occur.

4.2 Discussion & Suggestions for Further Studies

The estimators $\hat{\beta}$ coincides with the authors' expectation and the theory of economics.

The authors expected the Noncyclical Rate of Unemployment to be negative and the Employment Rate to be positive. The former, in contrast to the latter, would negatively impact GDP growth - that is, a drop in unemployment rate would correlate with the heightened level of productivity, thereby the factor if decreasing would make increments to the GDP.

Regarding the Currency Conversions (USD/EUR), the expected output is a positive estimator. In the short-term, the appreciation of USD might decrease the demand on products made in the U.S., thereby adversely impacting the U.S. GDP. Despite the fact, in the contemporary situation, the increased level of exchange rate in the United States has been considered as a result from an underlying shock, rather than the demand on products - that is, the widening of the monetary policy gap between the Federal Reserve and the ECB. In particular, the Federal Reserve started hiking interest rates in March; this was followed by further, faster hikes. The ECB, however, was slower to act, only announcing its first rate-hike in July 2022.

The obtained estimator $\hat{\beta}$ does not hold the assumption of Homoscedasticity and Non-autocorrelation. In this case, OLS estimators will still give unbiased (and also consistent) coefficient estimates. However, the estimators are inefficient - that is, they no longer have the minimum variance among the class of unbiased estimators (for the Homoscedasticity problem) and the standard error estimates could be wrong (for the autocorrelation problem). Hence, the estimators obtained by the

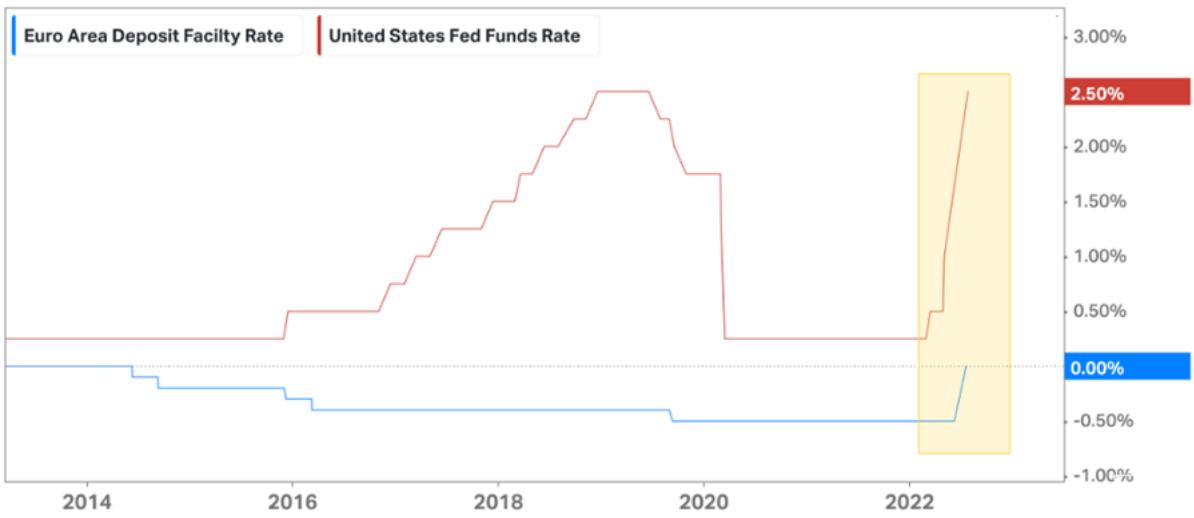

Figure 4.6: Interest rates in Europe and the U.S (Source: Capital.com, Data: Koyfin)

OLS method, in this case, can be considered as inefficient. For further studies, the method of Generalized Least Square (GLS) is suggested, in order to tackle Heteroskedasticity. For the autocorrelation problem, GLS is also a suggested remedy, but, as mentioned, the issue can be violated for time-series data, which is in this case. Therefore, GLS should be under careful consideration in further studies regarding the autocorrelation problem.

Chapter 5

Conclusion

This paper provided empirical analysis on the impact of several macroeconomic indicators on the Gross Domestic Product using the OLS method, and verifying the assumptions of the method on the obtained estimator. The authors also discussed the compatibility of the obtained results with the theories of economics and the contemporary situations in the U.S. The coefficients are demonstrated to be inefficient, with serious problems of Heteroskedasticity and autocorrelation. Thus, some further research is needed to study the movement of GDP.

Chapter 6

Teammate Evaluation

6.1 Ton Nu Trieu Man - MAMAIU20037

In this project, I would consider all members in this group to have equal contribution to this research project. Their roles can be demonstrated on the tasks assigned, the outcome's analysis and the dedication into working as a team. My team has been working with all members involving in all tasks, so my consideration is likely to based on their dedication. My evaluation regarding the performance of my teammates:

- Trong Tan has been a supportive teammate who devoted to both research and teamwork activities. He has assisted all other members and completed his tasks carefully and thoroughly. Our project can consider many new aspects thanks to his contribution regarding the socio-economic and statistical knowledge. For his huge contribution, I would definitely rate 5/5.

- Viet Hang is a wise member of our group. Her contribution has built up much of our team's strategy to study the real-world applications of the theories. Our work, thus, could proceed faster and discover many more aspects. She also put efforts into studying the problem and methods. Her contribution worth 5/5.

Regarding the efficiency of teamwork, we all realized that our work could go smoothly with participation of all members in the team. Each individual task, thereby, could be analyzed be the whole team, thus, become well-rounded. We also prioritized the effective communication among the team - that is, all updates about the project and progress are informed to all members, and we are all available to make changes if necessary.

6.2 Tran Viet Hang - MAMAIU18079

From my perspective, each member of my group contributed equally to this research project. It is interesting that each of us has different strengths and favorite fields of study that could help us

support each other well to accomplish this project.

I would like to give some words for my wonderful teammates:

- Trieu Man demonstrated a deep and thorough comprehension of the topic from the beginning. She handled tasks related to coding excellently thanks to her advanced skills in Python - a programming language. She was also diligent and hard-working to meet the tight deadline of the project. I would give her a 5/5.
- Trong Tan's efforts to keep the entire team informed of all the tasks were greatly appreciated. He was agile, supportive and always brings positive energy to lighten up our group's spirit whenever we had to face a problem arising from the collected data. I would give him a 5/5.

I think that my team worked well together to achieve the goal of this research project. Each individual took full responsibility for his/her task to complete it on time.

6.3 Le Trong Tan - MAMAIU20032

Personally speaking, members in this group contribute equally to this research project. Each member was assigned different tasks, for example: collecting and pre-processing data in Microsoft Excel and Python, executing the algorithm, and making reports and slides.

Regarding the performance of my teammates:

- Man has excelled in this project. She is extremely motivated and hard-working and displays a deep and thorough understanding of the research topic. She approaches the topic with critical thought and careful work and tries her best to figure out ways to deal with the complexity of data. I would give her a 5/5.
- Hang is a dedicated member of our group. She shows a good level of understanding of our topics and statistics concepts. Besides, she always helps others in the groups if needed. I would also give her a 5/5.

I believe that my team effectively collaborated to accomplish the objective of this research project. Everyone was responsible for their job and worked together to effectively fix any issues that could have arisen. We should update the members on the progress of the work often so that everyone is aware of what is happening, this is one thing I would change.

Bibliography

[1] Brooks, C., 2021. Introductory econometrics for finance. 3rd ed. [ebook] Available at: https://hk1lib.org/book/2338769/7b9834

[2] Callen, T. (2019) GROSS DOMESTIC PRODUCT: AN ECONOMY'S ALL. International Monetary Fund. Available at: https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/gross-domestic-product-GDP

[3] "Could the euro collapse?, Paid for and posted by Capital.com." 2022. Reuters. https://www.reuters.com/article/sponsored/could-the-euro-collapse

[4] "How currency appreciations affect growth, World Economic Forum." 2015. The World Economic Forum. https://www.weforum.org/agenda/2015/08/how-currency-appreciations-affect-growth/.

[5] Google Colaboratory https://colab.research.google.com/drive/1crh6zwKMnu4IGKHDQrBVP-sLcBDlVJ52?usp=sharing