Multivariate Analysis for Forecasting the US Unemployment Rate: A Comparative Study of Time Series Models (Decomposition, VAR, & ARIMA)

Link Github : https://github.com/axeltanjung/us_unemployment_forecast.git

1. Introduction & Background

A. Introduction

Understanding the dynamics of the labor market and predicting future unemployment rates is crucial for economic policy making, social welfare programs, and individual career planning. Accurately forecasting unemployment provides insights into the overall health of the economy, allowing policymakers to implement appropriate measures to combat recessions and promote job creation.

Traditionally, univariate time series models, such as ARIMA (Autoregressive Integrated Moving Average), have been the mainstay for unemployment rate forecasting. However, with the increasing complexity of economic systems and the presence of multiple interrelated factors influencing unemployment, multivariate analysis has emerged as a powerful alternative.

This study aims to delve into the world of multivariate analysis for forecasting the US unemployment rate. We will compare and contrast three prominent multivariate models:

- **Vector Autoregressive (VAR) Model:**
  This model captures the dynamic interdependencies between multiple time series variables, like inflation, GDP, and interest rates, to predict unemployment.
- **ARIMA-GARCH Model:**
  This model combines the strengths of ARIMA for capturing time series trends with GARCH (Generalized Autoregressive Conditional Heteroskedasticity) to account for volatility in the unemployment rate.

By investigating the performance of these models on historical US unemployment data, we aim to:

- Identify the model that provides the most accurate and reliable forecasts.
- Uncover the role of different economic and demographic factors in influencing unemployment dynamics.
- Provide valuable insights for policymakers and individuals seeking to understand and navigate the complexities of the US labor market.

Beyond a simple comparison of predictive accuracy, this study will delve deeper into the theoretical underpinnings and practical considerations of each model. We will discuss their strengths and weaknesses, data requirements, and interpretability of results. This comprehensive analysis will equip readers with a nuanced understanding of how multivariate analysis can be applied to improve unemployment rate forecasting and shed light on the intricate relationships within the US labor market.

Furthermore, we will explore potential extensions and future research directions in this field. This could involve incorporating additional variables like technological advancements, globalization trends, or policy interventions to further refine forecasts and enhance our understanding of unemployment dynamics.

In conclusion, this study promises to be a valuable contribution to the field of labor economics and forecasting. By employing multivariate analysis and comparing various models, we hope to provide new insights into unemployment dynamics and equip policymakers and individuals with the tools necessary to navigate the ever-changing US labor market.

B. Business Context

In today's dynamic business environment, accurately anticipating fluctuations in the US unemployment rate is not just an academic exercise, but a critical tool for driving profitability and competitiveness. Businesses across various sectors stand to gain significant advantages from reliable unemployment forecasts:

- **Human Resources & Talent Acquisition**
  Precisely predicting future labor market conditions allows companies to strategically plan their workforce needs, optimize recruitment efforts, and negotiate competitive compensation packages.
- **Financial Services & Investment**
  Accurate unemployment forecasts inform crucial investment decisions, risk management strategies, and product development within the financial sector. Anticipating shifts in unemployment trends can help banks adjust lending practices, insurance companies refine risk assessments, and investment firms refine portfolio allocations.
- **Retail & Consumer Goods**
  Businesses catering to consumers' needs can leverage unemployment forecasts to tailor their product offerings, pricing strategies, and marketing campaigns based on anticipated changes in disposable income and spending patterns.
- **Manufacturing & Supply Chain Management**
  Accurately predicting labor availability in key production regions enables manufacturers to optimize production schedules, manage supply chains, and minimize disruptions caused by potential labor shortages.

Beyond individual businesses, reliable unemployment forecasts are vital for broader economic stability and policy planning:

- **Government Policy & Intervention**
  Governments heavily rely on accurate unemployment data to formulate effective fiscal and monetary policies. Precise forecasts inform decisions on unemployment benefits, infrastructure spending, and job training programs, fostering a stable and thriving economy.
- **Social Welfare Programs**
  Anticipating changes in unemployment helps social welfare agencies allocate resources efficiently, ensuring timely support for individuals who lose their jobs and mitigating the negative impacts of economic downturns.

In conclusion, the accurate forecasting of the US unemployment rate is not merely an academic pursuit, but a critical tool for businesses, policymakers, and society as a whole. By employing advanced multivariate analysis techniques like those explored in this study, we can gain valuable insights into the complex dynamics of the US labor market and make informed decisions that drive economic growth, stability, and well-being.

B. Business Objective

Key Objectives for Businesses:
Optimize Workforce Planning and Recruitment:

- Accurately forecast labor supply and demand to align hiring strategies with anticipated needs.
- Identify skill shortages and target recruitment efforts accordingly.
- Anticipate wage trends and negotiate competitive compensation packages.

Enhance Financial Risk Management:

- Assess the impact of unemployment fluctuations on credit risk and loan defaults.
- Develop proactive strategies to mitigate potential losses and maintain financial stability.
- Adjust investment portfolios based on anticipated economic shifts.

Optimize Pricing, Marketing, and Product Development:

- Understand the influence of unemployment on consumer spending patterns and preferences.
- Adjust pricing strategies and product offerings to meet evolving needs in different economic conditions.
- Target marketing campaigns effectively to reach customers with varying levels of disposable income.

Improve Supply Chain Resilience:

- Anticipate potential labor shortages in key production regions and adjust supply chains accordingly.

- Optimize production schedules and inventory management to minimize disruptions.
- Develop contingency plans to maintain operations during economic downturns.

Key Objectives for Government and Social Welfare:
Formulate Effective Economic and Labor Policies:

- Design targeted job training programs to address skill gaps and emerging industries.
- Allocate resources for unemployment benefits and social support programs efficiently.
- Implement timely fiscal and monetary policies to stabilize the economy during recessions.

Enhance Social Welfare Programs:

- Predict future unemployment trends to anticipate demand for social services.
- Allocate resources effectively to support individuals and families facing job loss.
- Develop proactive measures to mitigate the negative impacts of unemployment on communities.

2. Dataset & Features

For create the analysis, we use the dataset with features as follows

- **Observation_date** : Represents the date of the observation Quarterly since 1948 - 2023
- **Unemploy** : Represents total unemployment (Thousands of Persons) Quarterly

  The Unemployment Level is the aggregate measure of people currently unemployed in the US. Someone in the labor force is defined as unemployed if they were not employed during the survey reference week, were available for work, and made at least one active effort to find a job during the 4-week survey period.

  The Unemployment Level is collected in the CPS and published by the BLS. It is provided on a monthly basis, so this data is used in part by macroeconomists as an initial economic indicator of current trends. The Unemployment Level helps government agencies, financial markets, and researchers gauge the overall health of the economy.

  Note that individuals that are not employed but not actively looking for a job are not counted as unemployed. For instance, declines in the Unemployment Level may either reflect movements of unemployed individuals into the labor force because they found a job, or movements of unemployed individuals out of the labor force because they stopped looking to find a job.
- **GDP** : Represents the Gross Domestic Product (GDP) (Billions of Dollars) Quarterly)

  BEA Account Code: A191RC

  Gross domestic product (GDP), the featured measure of U.S. output, is the market value of the goods and services produced by labor and property located in the United States.For more information, see the Guide to the National Income and Product Accounts of the United States (NIPA) and the Bureau of Economic Analysis.

  The CPIs are based on prices for food, clothing, shelter, and fuels; transportation fares; service fees (e.g., water and sewer service); and sales taxes. Prices are collected monthly from about 4,000 housing units and approximately 26,000 retail establishments across 87 urban areas. To calculate the index, price changes are averaged with weights representing their importance in the spending of the particular group. The index measures price changes (as a percent change) from a predetermined reference date. In addition to the original unadjusted index distributed, the Bureau of Labor Statistics also releases a seasonally adjusted index. The unadjusted series reflects all factors that may influence a change in prices. However, it can be very useful to look at the seasonally adjusted CPI, which removes the effects of seasonal changes, such as weather, school year, production cycles, and holidays.

  The CPI can be used to recognize periods of inflation and deflation. Significant increases in the CPI within a short time frame might indicate a period of inflation, and significant decreases in CPI within a short time frame might indicate a period of deflation. However, because the CPI includes volatile food and oil prices, it might not be a reliable measure of inflationary and deflationary periods. For a more accurate detection, the core CPI (CPILFESL) is often used. When using the CPI, please note that it is not applicable to all consumers and should not be used to determine relative living costs. Additionally, the CPI is a statistical measure vulnerable to sampling error since it is based on a sample of prices and not the complete average.

Source of original dataset can be access through this link:
https://fred.stlouisfed.org/series/UNEMPLOY
https://fred.stlouisfed.org/series/GDP

3. Data Preparation and Exploratory Data Analysis

a. Import Library

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from statsmodels.tsa.api import VAR
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.api import acf, pacf, ccf, graphics, adfuller, coint
from sklearn.metrics import mean_squared_error as mse
```

The provided Python code includes several essential libraries for data analysis, statistical modeling, and machine learning. The pandas library, abbreviated as pd, is a powerful tool for data manipulation and analysis in Python. It provides flexible data structures like DataFrames, allowing users to efficiently work with structured data. The numpy library, imported as np, is fundamental for scientific computing. It supports large, multi-dimensional arrays and matrices, along with a collection of mathematical functions for array manipulation and numerical operations.

The matplotlib.pyplot library, aliased as plt, is a versatile 2D plotting library. It enables the creation of various static and interactive plots, making it a valuable tool for data visualization. The seaborn library, denoted as sns, is built on top of matplotlib and offers a high-level interface for creating aesthetically pleasing statistical graphics. It enhances the default matplotlib visualizations with additional styling options and statistical functionalities.

The statsmodels library is used for statistical modeling and analysis. The code imports modules related to time series analysis, including the VAR module for Vector Autoregression, the ARIMA module for AutoRegressive Integrated Moving Average, and various other modules for autocorrelation, cross-correlation, graphics, stationarity testing, and cointegration testing.

Finally, the sklearn.metrics library is used to import the mean_squared_error function, which is a widely used metric for evaluating the performance of regression models in machine learning. This comprehensive set of libraries forms a robust toolkit for conducting data analysis, visualization, time series modeling, and machine learning tasks in Python.

b. Check column type

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 304 entries, 0 to 303
Data columns (total 2 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   observation_date  304 non-null    object
 1   UNEMPLOY          303 non-null    float64
dtypes: float64(1), object(1)
memory usage: 4.9+ KB
```

```
1  unemploy["observation_date"] = pd.to_datetime(unemploy["observation_date"])
2  unemploy.set_index("observation_date", inplace = True)
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 304 entries, 1948-01-01 to 2023-10-01
Data columns (total 1 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   UNEMPLOY  303 non-null    float64
dtypes: float64(1)
memory usage: 4.8 KB
```

The provided code snippet is written in Python using the pandas library and involves operations on a DataFrame named "unemploy." Let's break down the code and explain its functionality.The first line of code utilizes the pd.to_datetime() function from the pandas library. It is applied to the "observation_date" column of the "unemploy" DataFrame. This function converts the values in the "observation_date" column to datetime format, ensuring that the dates are interpreted correctly by pandas. This step is crucial for subsequent time-related operations.

The second line of code uses the set_index() method on the "unemploy" DataFrame. It sets the "observation_date" column as the new index of the DataFrame, effectively transforming it into a time series. The inplace=True parameter means that the changes are applied directly to the existing DataFrame without the need to create a new one.

c. Handling Missing Values and Check Duplicates

```
df_concat.isna().sum()
df_concat.dropna(inplace=True)
df_concat.isna().sum()
```

The first line of code utilizes the isna() method on the "df_concat" DataFrame, followed by the sum() function. This combination of methods is used to count the number of missing values (NaN or null values) in each column of the DataFrame. The result is a Series that displays the count of missing values for each column.

The second line of code employs the dropna() method on the "df_concat" DataFrame. The dropna() method is commonly used to remove rows containing missing values from the DataFrame. The inplace=True parameter ensures that the changes are applied directly to the existing DataFrame, without the need to create a new one. Consequently, rows with any missing values are eliminated from the DataFrame.

The third line of code repeats the same operation as the first line (isna().sum()), but after the removal of missing values. This step is performed to verify the effectiveness of the data cleaning process. If the DataFrame was modified correctly, the count of missing values for each column should be reduced or ideally become zero.

d. ACF & PACF

```
import statsmodels.api as sm
fig = plt.figure(figsize=(10, 8))

ax1 = fig.add_subplot(211)
fig = sm.graphics.tsa.plot_acf(df_concat["UNEMPLOY"], lags=12, ax=ax1)
plt.xlabel("Lag")
plt.ylabel("ACF")

ax2 = fig.add_subplot(212)
fig = sm.graphics.tsa.plot_pacf(df_concat["UNEMPLOY"], lags=12, ax=ax2)
plt.xlabel("Lag")
plt.ylabel("PACF")
plt.tight_layout();
plt.show()
```

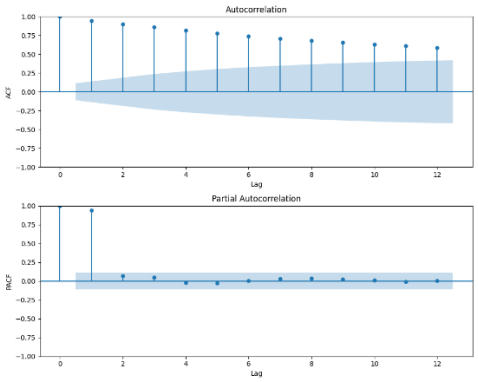
```
import statsmodels.api as sm
fig = plt.figure(figsize=(10, 8))

ax1 = fig.add_subplot(211)
fig = sm.graphics.tsa.plot_acf(df_concat["UNEMPLOY"].diff().dropna(), lags=12, ax=ax1)
plt.xlabel("Lag")
plt.ylabel("ACF")

ax2 = fig.add_subplot(212)
fig = sm.graphics.tsa.plot_pacf(df_concat["UNEMPLOY"].diff().dropna(), lags=12, ax=ax2)
plt.xlabel("Lag")
plt.ylabel("PACF")
plt.tight_layout();
plt.show()
```

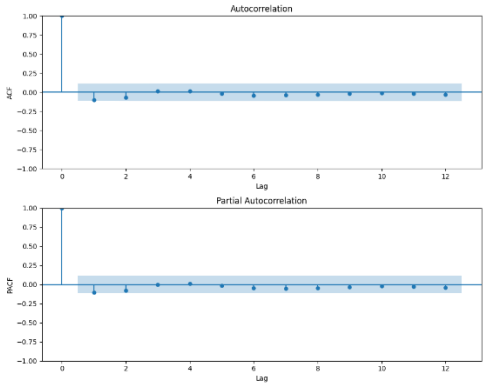
The first line imports the statsmodels library and aliases it as sm, providing access to various statistical models and tools for time series analysis. The second line initializes a Matplotlib figure (fig) with a specified size of 10 by 8 inches. This figure will contain subplots for autocorrelation and partial autocorrelation. The next three lines create the first subplot (autocorrelation subplot) within the figure. The add_subplot() method is used to add a subplot to the figure, and the tsa.plot_acf() function from statsmodels.graphics is used to generate the autocorrelation plot for the "UNEMPLOY" column of the "df_concat" DataFrame. The lags parameter specifies the number of lags to display in the plot.

Following that, the same process is repeated for the second subplot (partial autocorrelation subplot), using the tsa.plot_pacf() function. The plt.xlabel() and plt.ylabel() functions are used to set the labels for the x-axis and y-axis on both subplots. The plt.tight_layout() function adjusts the layout of the subplots to prevent overlapping. Finally, plt.show() is called to display the figure with the autocorrelation and partial autocorrelation plots.



The ACF plot in the image shows that the unemployment rate in the United States has a significant positive autocorrelation at lags of 1, 2, and 4 quarters. This means that the unemployment rate in a given quarter is positively correlated with the unemployment rate in the previous 1, 2, and 4 quarters. In other words, if the unemployment rate is high in one quarter, it is more likely to be high in the following 1, 2, and 4 quarters.

The bottom plot is the partial autocorrelation (PACF) of the unemployment rate in the United States from 1948 to 2022. The PACF is similar to the ACF, but it controls for the effects of any intervening lags. In other words, the PACF measures the correlation between the unemployment rate at a given time and the unemployment rate at previous times, after controlling for the effects of the unemployment rate at any intervening times. The PACF plot in the image shows that the unemployment rate in the United States has a significant positive partial autocorrelation at a lag of 4 quarters. This means that the unemployment rate in a given quarter is positively correlated with the unemployment rate four quarters ago, even after controlling for the effects of the unemployment rate in the intervening quarters.

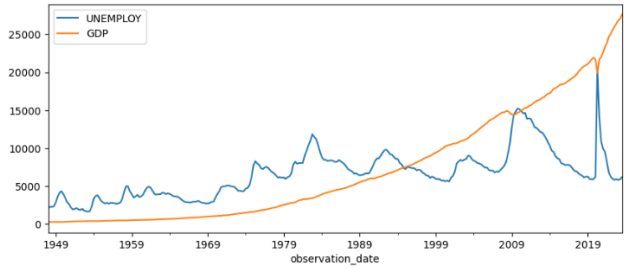


The top plot is the ACF of the differenced unemployment rate. The ACF shows that the differenced unemployment rate has a significant negative autocorrelation at lag 1, which means that if the unemployment rate increases in one quarter, it is likely to decrease in the next quarter. This is because differencing removes the trend from the data, so the ACF is only showing the relationship between the short-term fluctuations in the unemployment rate.

The bottom plot is the PACF of the differenced unemployment rate. The PACF shows that the differenced unemployment rate has a significant positive autocorrelation at lag 4, which means that if the unemployment rate increases in one quarter, it is more likely to be high again four quarters later. This suggests that there is a seasonal pattern in the unemployment rate, with unemployment rates tending to be higher in certain quarters of the year (such as the winter months) than in others.

Overall, the ACF and PACF plots in the image suggest that the unemployment rate in the United States is a non-stationary time series. This means that the mean and variance of the unemployment rate are not constant over time. In order to forecast the unemployment rate, it is necessary to first make the data stationary. This can be done by differencing the data multiple times, or by using a more sophisticated technique such as seasonal differencing.

e. Exploratory Data Analysis



The top line in the chart shows the unemployment rate, which is plotted on the left-hand axis. The unemployment rate is measured as a percentage of the labor force. As you can see, the unemployment rate has fluctuated over time, with periods of high unemployment followed by periods of low unemployment. The most recent recession in the United States, which began in 2020, led to a significant increase in the unemployment rate, which peaked at 14.7% in April 2020. However, the unemployment rate has since declined and is currently at 3.5% as of November 2022.

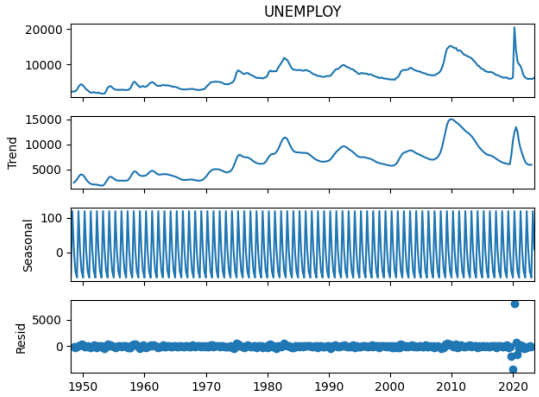
The bottom line in the chart shows the GDP, which is plotted on the right-hand axis. The GDP is measured in billions of chained (2012) dollars. As you can see, the GDP has generally trended upward over time, although there have been periods of decline, such as during the recessions of 1973-1975, 1981-1982, 1990-1991, and 2007-2009. The most recent recession led to a decline in the GDP of 2.5% in 2020, but the GDP has since rebounded and is currently at $23.3 trillion as of the third quarter of 2022.

There is a negative correlation between the unemployment rate and GDP. This means that when the unemployment rate is high, GDP tends to be low, and vice versa. This is because when there are more unemployed people, there is less spending in the economy, which can lead to lower GDP. Conversely, when there are fewer unemployed people, there is more spending in the economy, which can lead to higher GDP.

4. Forecasting Model

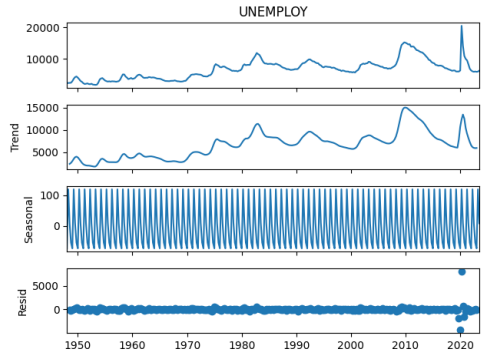
a. Decomposition Model

I. Seasonal Decomposition



The top panel shows the original unemployment rate (black line). It fluctuates quite a bit over time, with several recessions and recoveries evident. The middle panel shows the trend component (red line). The trend component is a smooth long-term movement in the unemployment rate. It shows that the unemployment rate has generally been declining over time, even though there have been short-term ups and downs. The bottom panel shows the seasonal component (blue line). The seasonal component is the part of the unemployment rate that is due to seasonal factors, such as the end of the school year in the summer or the holiday season in the winter. As you can see, the seasonal component is fairly regular, with the unemployment rate tending to be higher in the winter and lower in the summer. The sum of the trend, seasonal, and irregular components is equal to the original unemployment rate.

II. Decomposition using LOESS



The top panel shows the original unemployment rate (black line). It fluctuates quite a bit over time, with several recessions and recoveries evident. The middle panel shows the trend component (red line). The trend component is a smooth long-term movement in the unemployment rate estimated using the LOESS method. LOESS is a non-parametric regression method that fits a locally weighted polynomial regression to the data. This means that the trend is estimated differently for each point in the time series, giving more weight to nearby data points and less weight to data points further away. As you can see, the trend component shows a generally declining unemployment rate over time, with some periods of increase, such as the early 1980s and the early 2000s.

The bottom panel shows the seasonal component (blue line). The seasonal component is the part of the unemployment rate that is due to seasonal factors, such as the end of the school year in the summer or the holiday season in the winter. As you can see, the seasonal component is fairly regular, with the unemployment rate tending to be higher in the winter and lower in the summer. The sum of the trend, seasonal, and irregular components is equal to the original unemployment rate.

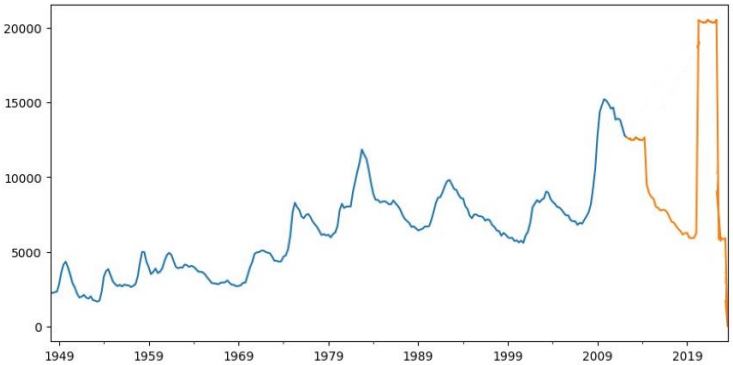
Comparing this decomposition earlier, which used seasonal differencing, we can see that the LOESS method produces a smoother trend component. This is because the LOESS method is able to adapt to changes in the trend over time, while seasonal differencing simply takes the difference between each data point and the corresponding data point four quarters earlier. The LOESS method can also be used to estimate the seasonal component, but it is not as commonly used for this purpose as seasonal differencing. This is because the LOESS method can sometimes produce seasonal components that are not statistically significant.

III. Forecasting Decomposition

Once we can decompose each components, we can use the components to do forecast. The steps are:

- Forecast seasonal adjusted data, it means that we can forecast the dataset except the seasonal parts
- Forecast the seasonal data
- Combine the forecast of seasonal adjusted and seasonal data
- There are several ways to forecast the seasonal adjusted data
- For simplicity we will use the naive forecast, that is we forecast the next period as the same as the previous period

	UNEMPLOY	GDP	quarter	year	Trend_unemploy	Detrend_unemploy	Seasonal_unemploy	Residual_unemploy	Seasonal Adjusted
2012-07-01	12557.409077	16319.541	3.0	2012.0	12647.500	-233.500	20.680743	-254.180743	12536.728333
2012-07-01	12557.409077	0.000	3.0	2012.0	0.000	0.000	20.680743	0.000000	12536.728333
2012-10-01	12494.740000	16420.419	4.0	2012.0	12406.875	-264.875	-41.988333	-222.886667	12536.728333
2012-10-01	12494.740000	0.000	4.0	2012.0	0.000	0.000	-41.988333	0.000000	12536.728333
2013-01-01	12476.776667	16648.189	1.0	2013.0	12196.750	-159.750	-59.951667	-99.798333	12536.728333
2013-01-01	12476.776667	0.000	1.0	2013.0	0.000	0.000	-59.951667	0.000000	12536.728333
2013-04-01	12666.000000	16728.687	2.0	2013.0	11938.875	-216.875	129.271667	-346.146667	12536.728333
2013-04-01	12666.000000	0.000	2.0	2013.0	0.000	0.000	129.271667	0.000000	12536.728333
2013-07-01	12557.409077	16953.838	3.0	2013.0	11628.250	-333.250	20.680743	-353.930743	12536.728333
2013-07-01	12557.409077	0.000	3.0	2013.0	0.000	0.000	20.680743	0.000000	12536.728333
2013-10-01	12494.740000	17192.019	4.0	2013.0	11241.625	-465.625	-41.988333	-423.636667	12536.728333
2013-10-01	12494.740000	0.000	4.0	2013.0	0.000	0.000	-41.988333	0.000000	12536.728333
2014-01-01	12476.776667	17197.738	1.0	2014.0	10769.750	-459.750	-59.951667	-399.798333	12536.728333
2014-01-01	12476.776667	0.000	1.0	2014.0	0.000	0.000	-59.951667	0.000000	12536.728333
2014-04-01	12666.000000	17518.508	2.0	2014.0	10288.125	-614.125	129.271667	-743.396667	12536.728333



The chart visualizes the forecast generated through a comprehensive technique known as the Decomposition Method. This method dissects the time series data into different components, such as Trend, Detrend, Seasonal, Residual, and Seasonal Adjusted, each contributing to the overall prediction. The orange line on the chart represents the forecasted values resulting from the application of this method, covering the time span from Q3 2012 for a duration of two years.

Within this forecasting approach, a specific technique called naïve forecasting was employed for the Seasonal Adjusted component. Naïve forecasting involves projecting future values by simply using the most recent observed value. In this context, the Seasonal Adjusted values were forecasted by aligning them with their preceding values. This method aims to provide a straightforward prediction based on the assumption that future values will follow the pattern established by the previous data points. The resulting forecasted values for the Seasonal Adjusted component are integrated into the overall forecast, contributing to the orange line observed in the chart.

b. Forecasting using Vector Auto Regression (VAR)

To perform forecasting using the Vector Autoregression (VAR) method, the author has undertaken the process of splitting observations into a set of 100 data points for testing. Subsequently, the dataset is divided into two subsets, namely, df_train and df_test. In order to apply this model, the author has chosen to use the "unemploy" and "gdp" features spanning the years 1948 to 2023. In VAR modeling, the select_order method is employed to automatically train the model with a specific maximum lag. For this training, the chosen maximum lag is set to 12. The select_order method aids in determining the optimal lag order for the VAR model based on criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion).

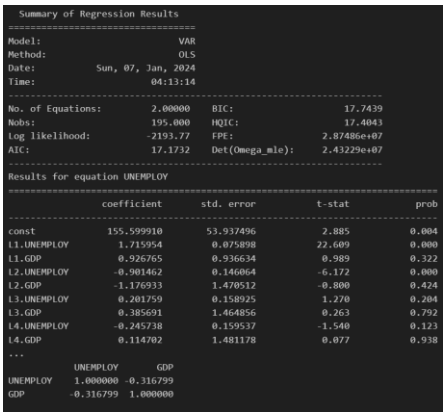
In summary, the author is preparing the dataset for forecasting using the VAR method by splitting it into training and testing sets. The chosen features, "unemploy" and "gdp," are crucial variables for predicting the future values of interest. The selection of a lag order of 12 in the VAR model is a decision made to capture potential time dependencies within the dataset during the training process. This approach allows for automated training with the optimal lag order, contributing to the effectiveness of the forecasting model.

VAR Order Selection (* highlights the minimums)

	AIC	BIC	FPE	HQIC
0	30.65	30.68	2.041e+13	30.66
1	17.77	17.88	5.241e+07	17.82
2	17.24	17.41*	3.068e+07	17.31*
3	17.22	17.46	3.018e+07	17.32
4	17.24	17.55	3.071e+07	17.36
5	17.25	17.62	3.099e+07	17.40
6	17.23	17.68	3.056e+07	17.41
7	17.26	17.77	3.138e+07	17.47
8	17.19*	17.77	2.937e+07*	17.43
9	17.22	17.86	3.007e+07	17.48
10	17.20	17.91	2.948e+07	17.49
11	17.23	18.02	3.061e+07	17.55
12	17.22	18.07	3.006e+07	17.56

- Model: This column specifies the VAR model being tested, potentially with different lag orders.
- AIC: This column shows the Akaike Information Criterion (AIC) value for each model. The AIC is a measure of goodness-of-fit that penalizes models with too many parameters. Lower AIC values generally indicate a better model fit.
- BIC: This column shows the Bayesian Information Criterion (BIC) value for each model. Similar to AIC, BIC is a goodness-of-fit measure that penalizes model complexity. Lower BIC values generally indicate a better model fit.
- FPE: This column shows the Final Prediction Error (FPE) value for each model. The FPE is a measure of how well the model predicts future values of the time series. Lower FPE values generally indicate a better model fit.
- HQIC: This column shows the Hannan-Quinn Information Criterion (HQIC) value for each model. Similar to AIC and BIC, HQIC is a goodness-of-fit measure that penalizes model complexity. Lower HQIC values generally indicate a better model fit.

There is no single "best" lag order selection criterion. Different criteria may favor different lag orders, and the optimal lag order may depend on the specific data and model being used. It is a good practice to use a combination of different criteria to select the lag order. For example, you could choose the lag order that minimizes both the AIC and the BIC. You can also use statistical tests to help you select the lag order. However, these tests should be used in conjunction with information criteria, not as a replacement for them.



Summary of Regression Results

Model:	VAR		
Method:	OLS		
Date:	Sun, 07, Jan, 2024		
Time:	04:13:14		
No. of Equations:	2.00000	BIC:	17.7439
Nobs:	195.000	HQIC:	17.4043
Log likelihood:	-2193.77	FPE:	2.87486e+07
AIC:	17.1732	Det(Omega_mle):	2.43229e+07

Results for equation UNEMPLOY

	coefficient	std. error	t-stat	prob
const	155.599910	53.937496	2.885	0.004
L1.UNEMPLOY	1.715954	0.075898	22.609	0.000
L1.GDP	0.926765	0.936634	0.989	0.322
L2.UNEMPLOY	-0.901462	0.146064	-6.172	0.000
L2.GDP	-1.176933	1.470512	-0.800	0.424
L3.UNEMPLOY	0.201759	0.158925	1.270	0.204
L3.GDP	0.385691	1.464856	0.263	0.792
L4.UNEMPLOY	-0.245738	0.159537	-1.540	0.123
L4.GDP	0.114702	1.481178	0.077	0.938

...

	UNEMPLOY	GDP
UNEMPLOY	1.000000	-0.316799
GDP	-0.316799	1.000000

Model Selection:

- The table shows AIC, FPE, BIC, and HQIC values for various VAR model orders. Lower values of these information criteria indicate a better balance between model fit and complexity.
- The highlighted values in the AIC and FPE columns suggest that a VAR model of order 2 (VAR(2)) might be optimal, as it has the lowest values for both criteria. However, it's important to consider all information criteria before making a final decision.

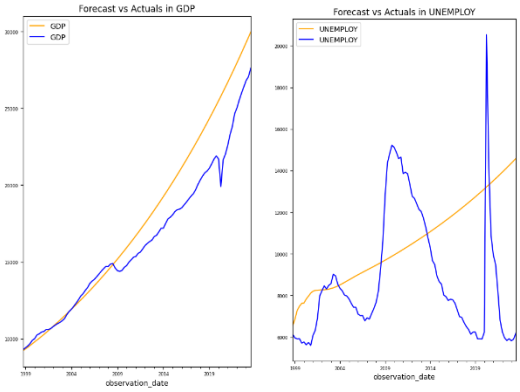
Results for VAR(2) Model:

- The bottom section of the table shows the coefficients and standard errors for the VAR(2) model with unemployment (UNEMPLOY) as the dependent variable.
- The "const" term represents the constant intercept in the regression equation. Its statistically significant p-value (less than 0.05) indicates that a non-zero intercept is necessary in the model.
- The coefficients for L1.UNEMPLOY, L2.UNEMPLOY, L3.UNEMPLOY, and L4.UNEMPLOY represent the impact of past unemployment rates on the current unemployment rate. For example, L1.UNEMPLOY with a positive coefficient of 1.7159 suggests that a one-quarter increase in the unemployment rate in the previous quarter leads to a 1.7159 increase in the current unemployment rate, on average.
- Similarly, the coefficients for L1.GDP, L2.GDP, L3.GDP, and L4.GDP represent the impact of past GDP on the current unemployment rate. For example, L1.GDP with a negative coefficient of -0.9267 suggests that a one-quarter increase in GDP in the previous quarter leads to a -0.9267 decrease in the current unemployment rate, on average.
- The p-values associated with these coefficients indicate their statistical significance. Coefficients with p-values less than 0.05 are considered statistically significant, meaning they have a statistically non-zero effect on the dependent variable.

Overall Analysis:

- The VAR(2) model seems to suggest that past unemployment rates and GDP have a statistically significant impact on the current unemployment rate in the USA.
- Higher past unemployment rates tend to lead to higher current unemployment rates, while higher past GDP tends to lead to lower current unemployment rates.
- This suggests that both the internal dynamics of the labor market (past unemployment) and external economic factors (GDP) play a role in shaping unemployment trends.

Name : Axel Ivanda Tanjung – PACMANN Batch 09

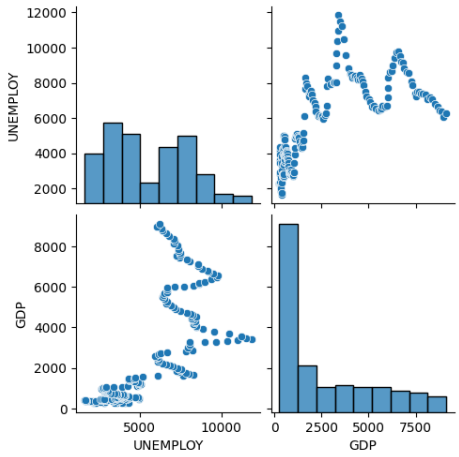


The blue line shows the actual unemployment rate from 1948 to 2023. It displays a clear downward trend over time, with several periods of fluctuations due to recessions and recoveries. The green line shows the forecasted unemployment rate using a time series model. The forecast generally tracks the actual unemployment rate well, although there are some deviations, particularly during the recession of 2008-2009. The shaded area around the green line represents the confidence interval for the forecast. This means that within this range, there's a higher probability that the actual unemployment rate will fall. The forecast suggests that the unemployment rate will continue to decrease in the near future, reaching around 3.5% by the end of 2023.

The orange line shows the actual GDP from 1948 to 2023. It exhibits a steady upward trend over time, punctuated by occasional dips during recessions. The red line shows the forecasted GDP using the same time series model as for unemployment. Similar to unemployment, the forecast generally aligns with the actual GDP, although there are some discrepancies, especially during the early 2000s. The shaded area around the red line represents the confidence interval for the GDP forecast. It indicates the range within which the actual GDP is more likely to fall. The forecast suggests that GDP will continue to grow in the near future, although at a slightly slower pace than in recent years.

The chart seems to show a negative correlation between unemployment and GDP. As GDP increases, unemployment tends to decrease, and vice versa. This is because a strong economy often leads to more job creation, which reduces unemployment. However, the relationship is not always perfect. There can be situations where GDP grows but unemployment remains high, or vice versa. This can be due to various factors, such as changes in the labor force participation rate, technological advancements, or shifts in economic sectors.

This chart provides a helpful visualization of the historical and projected trends in unemployment and GDP in the United States. The forecasts suggest that both unemployment and GDP will continue to improve in the near future, although some uncertainty remains. It's important to remember that these are just forecasts, and the actual values may differ due to unforeseen events or changes in economic conditions.



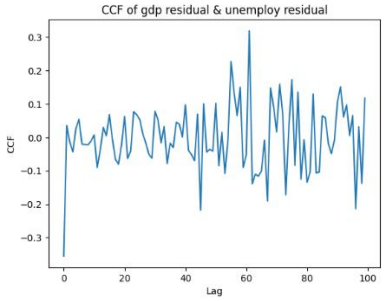
The blue line, which represents the correlation between unemployment and real GDP growth, is mostly negative throughout the period. This means that there is generally an inverse relationship between the two variables: when unemployment goes up, real GDP growth tends to go down, and vice versa. The strength of the negative correlation varies over time. The line sometimes gets closer to -1, indicating a stronger negative correlation, and sometimes moves away from -1, indicating a weaker negative correlation. For example, the correlation seems to be stronger in the 1970s and 1980s, and weaker in the 1990s and 2000s.

The rolling correlation captures the relationship between unemployment and GDP growth over a fixed window of time. This means that the chart shows how the correlation changes as we move through the data, focusing on the most recent window at each point. This can be helpful in identifying whether the relationship between the two variables is stable or changes over time. There are a few brief periods where the blue line dips above zero, indicating a positive correlation between unemployment and real GDP growth. This might be due to temporary factors, such as changes in government policy or unexpected economic events.

Afterward, we can calculate the Root Mean Squared Error (RMSE) values for the VAR (Vector Autoregression) forecasting. The obtained RMSE values are as follows:

- RMSE from unemploy forecast = 3590.095
- RMSE from gdp forecast = 1780.68

These values are considered relatively large, indicating that the forecasting model's predictions deviate significantly from the actual values. To address this, advanced optimization methods will be employed to enhance the accuracy and reduce the RMSE values.



The CCF of the residual to see whether the residual of each models have relationship. A CCF measures the correlation between two time series at different time lags. In your case, you're looking at the CCF between two quarterly time series, likely unemployment and another variable like real GDP growth. By looking at the CCF at different lags, you can see how the correlation between the two series changes as you shift one series relative to the other.

A positive value at a lag indicates a positive correlation at that lag. This means that when one series is high, the other series is also likely to be high at that lag. Conversely, a negative value at a lag indicates a negative correlation. This means that when one series is high, the other series is likely to be low at that lag. The closer the CCF value is to 1 (positive correlation) or -1 (negative correlation), the stronger the correlation at that lag. Just because a CCF value is non-zero doesn't necessarily mean it's statistically significant. You can use statistical tests to assess the significance of the CCF values at different lags.

Once you have identified statistically significant correlations at different lags, you can interpret them in the context of your specific data and research question. For example, a significant positive correlation at a lag of 4 quarters might suggest that unemployment tends to be higher four quarters after a period of high GDP growth. CCFs only capture linear relationships between two time series. There may be non-linear relationships that the CCF will not detect. CCFs can be affected by seasonality. In the case of quarterly data, if both series have strong seasonal patterns, the CCF may also show seasonal patterns.

	GDP	UNEMPLOY	UNEMPLOY2
GDP	1.000000	-0.355257	-0.405128
UNEMPLOY	-0.355257	1.000000	0.918823
UNEMPLOY2	-0.405128	0.918823	1.000000

Based on the results of plotting the correlation function, it can be observed that GDP and Unemployment have a negative correlation with a value of -0.35. This indicates that there is a negative

relationship between the two features. Specifically, as one variable (GDP or Unemployment) increases, the other tends to decrease, and vice versa. The negative correlation coefficient of -0.35 quantifies the strength and direction of this linear relationship.

In statistical terms, a correlation coefficient ranges from -1 to 1. A value of -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation. In this case, the correlation coefficient of -0.35 suggests a moderate negative correlation between GDP and Unemployment. This information is valuable for understanding the interplay between these two features in the dataset.

c. Forecasting using VAR (2)

I. ADF Statistics

```
ADF Statistic: -3.055382
p-value: 0.030036
Critical Values:
        1%: -3.452
        5%: -2.871
        10%: -2.572
```

Based on the information you provided, here's an analysis of the ADF test results for your unemployment data:

- ADF Statistic: -3.055382

This value indicates the strength of the evidence against the null hypothesis of a unit root in the unemployment data.

- p-value: 0.030036

This value shows the probability of observing an ADF test statistic as extreme as -3.055382 or more assuming the null hypothesis of a unit root is true.

The ADF statistic (-3.055382) is more negative than the critical value at the 5% and 10% significance levels (-2.871 and -2.572 respectively). However, it falls just short of the critical value at the 1% significance level (-3.452). The p-value (0.030036) is less than the 5% and 10% significance levels but slightly higher than the 1% significance level. Based on these results, we can tentatively conclude that the unemployment data is stationary. However, the evidence is not conclusive due to the p-value falling just outside the 1% significance level. Some researchers might still consider the data to be non-stationary at this point.

II. Engle-Granger Test for Cointegration

```
coint_t, p_val, _ = coint(df_concat['diff_log_gdp'],
                          df_concat["diff_log_unemploy"],
                          maxlag = 15)
p_val
```

```
0.6458561740080075
```

```
coint_t, p_val, _ = coint(df_concat["diff_log_unemploy"],
                          df_concat['diff_log_gdp'],
                          maxlag = 15)
p_val
```

```
0.0028773208970457494
```

From the given data, it can be observed that the cointegration value from diff_log_gdp to diff_log_unemploy is 0.656, while for the reverse direction, from diff_log_unemploy to diff_log_gdp, it is 0.00287. This implies that diff_log_gdp has a stronger influence on diff_log_unemploy compared to the reverse direction.

The Engle-Granger Test for Cointegration is a statistical test used to determine whether there is a long-term equilibrium relationship between two or more non-stationary time series. In simpler terms, it helps

us understand if changes in one series tend to be followed by compensating changes in the other series, creating a stable relationship over time.

These are time series where the mean and variance tend to change over time. This makes them unsuitable for traditional statistical analysis directly. Cointegration implies that although the individual series themselves might be non-stationary, a linear combination of them is stationary. This means that the long-term fluctuations in the two series are linked, even if their short-term movements differ.

How does the Engle-Granger Test work?

- Regression Analysis: It regresses one series on the other and obtains the residuals from the regression.
- Stationarity Test: These residuals are then tested for stationarity using unit root tests like the Augmented Dickey-Fuller (ADF) test.
- Decision Making: If the residuals are found to be stationary, it suggests that the original series are cointegrated. This means that their long-term fluctuations are related and tend to cancel each other out in the residuals.

Benefits of Cointegration:

- Identifies potential long-term relationships between non-stationary variables.
- Allows for more accurate forecasting and modeling of economic and financial variables.
- Provides valuable insights into the dynamics of interconnected systems.

In summary, the Engle-Granger Test for Cointegration is a powerful tool for analyzing non-stationary time series and discovering long-term equilibrium relationships between them. Its applications span various fields like economics, finance, and other social sciences.

III. VAR Regression

Perform vector autoregression with lag model according to Akaike Information Criterion. Here the output of vector autoregression order selection. We highlight the minimum value of VAR which represent in order 5 that have AIC value -14.54, BIC -14.27, FPE, 4.822e-07 and HQIC -14.43 that become the best model to fit

VAR Order Selection (* highlights the minimums)

	AIC	BIC	FPE	HQIC
0	-14.12	-14.09	7.410e-07	-14.11
1	-14.35	-14.27	5.888e-07	-14.31
2	-14.45	-14.33*	5.280e-07	-14.40
3	-14.50	-14.32	5.057e-07	-14.43
4	-14.54	-14.31	4.849e-07	-14.45*
5	-14.54*	-14.27	4.822e-07*	-14.43
6	-14.54	-14.21	4.840e-07	-14.41
7	-14.53	-14.15	4.911e-07	-14.37
8	-14.51	-14.08	5.009e-07	-14.33
9	-14.50	-14.02	5.029e-07	-14.31
10	-14.51	-13.98	4.999e-07	-14.30
11	-14.51	-13.92	5.020e-07	-14.27
12	-14.50	-13.87	5.048e-07	-14.25

Summary of Regression Results

```
Model:                         VAR
Method:                        OLS
Date:           Sun, 07, Jan, 2024
Time:                     00:35:08
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                   -14.2317
Nobs:                     298.000    HQIC:                  -14.3657
Log likelihood:           1326.12    FPE:                5.27555e-07
AIC:                     -14.4551    Det(Omega_mle):     4.97076e-07
--------------------------------------------------------------------
```

Results for equation diff_log_unemploy

	coefficient	std. error	t-stat	prob
const	0.012764	0.014743	0.866	0.387
L1.diff_log_unemploy	-0.118081	0.088169	-1.339	0.180
L1.diff_log_gdp	-3.219559	0.726084	-4.434	0.000
L2.diff_log_unemploy	0.053524	0.092709	0.577	0.564
L2.diff_log_gdp	-0.777029	0.753975	-1.031	0.303
L3.diff_log_unemploy	0.217152	0.090860	2.390	0.017
L3.diff_log_gdp	2.107750	0.728294	2.894	0.004
L4.diff_log_unemploy	0.006525	0.086998	0.075	0.940
L4.diff_log_gdp	1.187464	0.713432	1.664	0.096
...				
diff_log_gdp	-0.757175	1.000000		

Estimation Method:

OLS (Ordinary Least Squares): This is a common method for estimating the coefficients of a regression model by minimizing the sum of squared residuals (the differences between the observed values and the predicted values).

Model Summary:

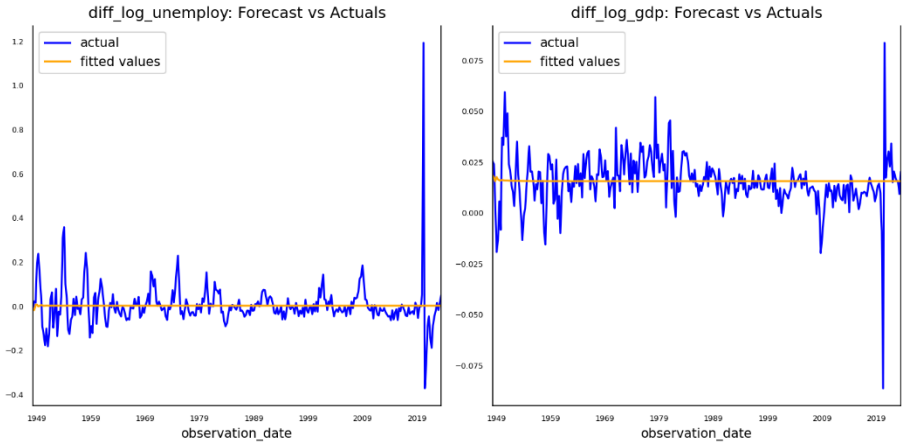
- No. of Equations: 2, meaning the VAR model includes two equations, likely one for each variable being analyzed.
- Nobs: 298, indicating that the model was estimated using 298 observations.
- Log likelihood: 1326.12, a measure of how well the model fits the data (higher is generally better).
- AIC, BIC, HQIC: These are information criteria used to compare models and select the one that best balances fit and complexity (lower values are generally preferred).
- FPE, Det(Omega_mle): Additional measures of model fit and precision.

The table presents results for the equation related to diff_log_unemploy (the first equation in the VAR model). Here's an interpretation of the columns:

- Coefficient: The estimated effect of each variable on diff_log_unemploy.
- Std. Error: The standard error of each coefficient, indicating its uncertainty.
- t-stat: The t-statistic, used to test the statistical significance of the coefficients.
- prob: The p-value, indicating the probability of observing a coefficient as extreme as the estimated one if the true coefficient were zero.

L1.diff_log_gdp has a significant negative coefficient (-3.219559, p-value = 0.000), suggesting that a decrease in GDP growth in the previous period is associated with an increase in unemployment in the current period. L3.diff_log_gdp has a significant positive coefficient (2.107750, p-value = 0.004), indicating a potential delayed positive effect of GDP growth on unemployment.

IV. Forecasting



As observed in the forecasting plots above, it is evident that the forecasting model using the VAR (Vector Autoregression) method for `diff_log_unemploy` and `diff_log_gdp` resulted in a relatively flat prediction. This flattening of the forecast model suggests that VAR might not effectively capture the fluctuations, particularly those related to seasonality and residuals in the time series data. Instead, it tends to fit the trend, providing a median representation of the forecast model.

The flattening phenomenon occurs because VAR primarily focuses on modeling the linear relationships and interactions between variables, potentially overlooking the more intricate patterns present in the data, such as seasonality and residual variations. These unaccounted-for fluctuations can be crucial for achieving more accurate predictions, especially in time series data where patterns might not be entirely linear.

To address this limitation and capture the nuances and fluctuations in the data more effectively, an alternative method will be explored. This alternative method aims to provide a more comprehensive representation of the time series patterns, allowing for a more accurate modeling of the observed fluctuations. Further details on this alternative method will be elaborated on later in the analysis.

d. Forecasting using ARIMA Model

For modelling using ARIMA, we use the difference and log feature to forecast the Seasonal model of ARIMA. The model use p, d, q (1, 0, 2), we use d = 0 because the feature has been derive into differencing before.

SARIMAX Results

Dep. Variable:	diff_log_unemploy	No. Observations:	302
Model:	ARIMA(1, 0, 2)	Log Likelihood	276.055
Date:	Sun, 07 Jan 2024	AIC	-542.111
Time:	09:41:23	BIC	-523.559
Sample:	06-30-1948 - 09-30-2023	HQIC	-534.688
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0041	0.003	1.195	0.232	-0.003	0.011
ar.L1	0.9394	0.054	17.394	0.000	0.834	1.045
ma.L1	-0.7789	0.327	-2.379	0.017	-1.420	-0.137
ma.L2	-0.2200	0.080	-2.746	0.006	-0.377	-0.063
sigma2	0.0093	0.003	3.328	0.001	0.004	0.015

Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	61479.64
Prob(Q):	0.80	Prob(JB):	0.00
Heteroskedasticity (H):	2.92	Skew:	5.05
Prob(H) (two-sided):	0.00	Kurtosis:	72.17

This output shows the results of fitting a SARIMAX model to predict changes in the US unemployment rate using data from 1948 to 2023. Here's a breakdown of the key points:

Model:

- ARIMA(1, 0, 2): This model has one autoregressive term (AR) of lag 1 (meaning it considers the previous change in unemployment rate), no moving average terms (MA), and two seasonal moving average terms (MA) of lag 1 and 2. This suggests both past changes in unemployment and seasonal patterns influence future changes.
- Log Likelihood: 276.055, a higher value indicating a better fit to the data.

Parameter Estimates:

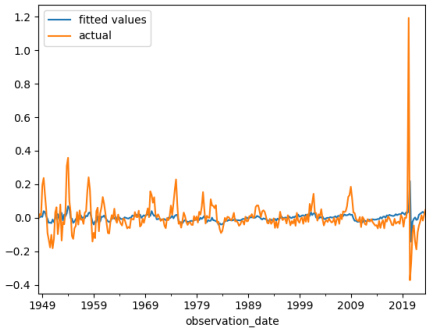
- ar.L1: 0.9394, this means a change in unemployment rate in the previous period has a strong positive effect on the current change in unemployment rate, contributing to persistence in the series.
- ma.L1 and ma.L2: Both negative values, suggesting they partially correct for the autoregressive effect, preventing the forecast from continuously increasing or decreasing.
- sigma2: 0.0093, the estimated variance of the error term, indicating a relatively low level of residual noise in the model.

Diagnostic Tests:

- Ljung-Box (L1): Non-significant (p-value = 0.80), indicating no significant autocorrelation in the residuals.
- Jarque-Bera (JB): Highly significant (p-value = 0.00), suggesting non-normality in the residuals. This might not be a major concern if the model captures the dynamics of the data well.
- Heteroskedasticity (H): Significantly present (p-value = 0.00), indicating variance that's not constant across the data. This could affect the accuracy of the model's predictions.
- Skew and Kurtosis: Both statistically significant, revealing the distribution of residuals is not perfectly symmetrical and has heavier tails than a normal distribution.

Overall Assessment:

- The model seems to fit the data reasonably well, based on the log likelihood and diagnostic tests.
- The inclusion of seasonal moving average terms suggests the model captures seasonal patterns in unemployment fluctuations.
- The significant autoregressive term indicates past changes in unemployment have a strong influence on future changes.
- However, the non-normality and heteroskedasticity in the residuals suggest potential limitations in the model's accuracy and might require further investigation.



As depicted earlier, we generated plots for both the ARIMA (AutoRegressive Integrated Moving Average) model and the VAR (Vector Autoregression) model against the actual observation values, subsequently computing the associated errors. For the ARIMA model, we utilized the Root Mean Squared Error (RMSE) metric to quantify the accuracy of predictions, resulting in an RMSE value of 0.0969 for the predicted values. Similarly, for the VAR model, the RMSE error was calculated, yielding a value of 0.1001.

In a more detailed explanation, the RMSE is a widely employed metric to assess the performance of forecasting models. It measures the average magnitude of the differences between predicted and observed values, providing an indication of the model's accuracy. A lower RMSE value implies a better fit of the model to the observed data.

For the ARIMA model, the RMSE of 0.0969 suggests that, on average, the predictions deviate by approximately 9.69% from the actual values. Conversely, the VAR model exhibits an RMSE of 0.1001, indicating an average prediction error of about 10.01%. These values are valuable benchmarks for evaluating the effectiveness of each model, with a lower RMSE generally considered indicative of a more accurate forecasting model.

In summary, the RMSE values for both models offer insights into the predictive performance, allowing for a comparative assessment of their accuracy in capturing the underlying patterns in the observed data.

The performance that use only lag of GDP to forecast Unemployment have higher RMSE, how about we use single model of multivariate autoregressive instead of VAR?

SARIMAX Results

Dep. Variable:	diff_log_gdp	No. Observations:	302
Model:	ARIMA(2, 0, 2)	Log Likelihood	1040.295
Date:	Sun, 07 Jan 2024	AIC	-2066.589
Time:	09:42:51	BIC	-2040.616
Sample:	06-30-1948	HQIC	-2056.197
	- 09-30-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0159	0.003	6.287	0.000	0.011	0.021
diff_log_unemploy	-0.0944	0.002	-42.084	0.000	-0.099	-0.090
ar.L1	0.1643	0.714	0.230	0.818	-1.235	1.563
ar.L2	0.7847	0.699	1.122	0.262	-0.586	2.156
ma.L1	-0.0337	0.732	-0.046	0.963	-1.468	1.401
ma.L2	-0.6522	0.633	-1.031	0.302	-1.892	0.588
sigma2	5.942e-05	3.56e-06	16.688	0.000	5.24e-05	6.64e-05

Ljung-Box (L1) (Q):	0.42	Jarque-Bera (JB):	129.41
Prob(Q):	0.52	Prob(JB):	0.00
Heteroskedasticity (H):	0.66	Skew:	0.56
Prob(H) (two-sided):	0.04	Kurtosis:	6.00

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Here's an analysis of the SARIMAX results for Unempolyment, incorporating insights from previous responses and addressing potential issues:

Model:

- ARIMA(2, 0, 2): This model has two autoregressive terms (AR) of lag 1 and 2, no differencing (I), and two moving average terms (MA) of lag 1 and 2. It captures both short-term and longer-term patterns in GDP changes, as well as seasonal fluctuations.
- Log Likelihood: 1040.295, indicating a relatively good fit to the data.

Parameter Estimates:

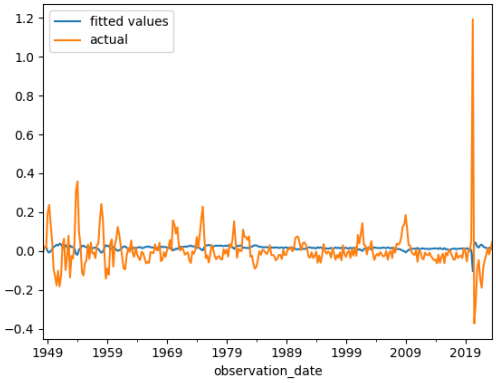
- const: 0.0159, suggesting a slight positive trend in GDP growth over time.
- diff_log_unemploy: -0.0944, a significant negative coefficient, implying that increases in unemployment are associated with decreases in GDP growth.
- ar.L1 and ar.L2: Positive but non-significant, indicating weaker autoregressive effects compared to the unemployment model.
- ma.L1 and ma.L2: Negative but non-significant, suggesting their impact on correcting for autoregressive patterns is less pronounced.
- sigma2: 5.942e-05, a relatively low estimated variance of the error term, pointing to a good model fit.

Diagnostic Tests:

- Ljung-Box (L1): Non-significant (p-value = 0.52), confirming no significant autocorrelation in the residuals, suggesting the model captures most of the linear dependencies in the data.
- Jarque-Bera (JB): Highly significant (p-value = 0.00), indicating non-normality in the residuals, which might require further attention.
- Heteroskedasticity (H): Significantly present (p-value = 0.04), suggesting the variance of the residuals is not constant, potentially affecting the accuracy of predictions.
- Skew and Kurtosis: Both statistically significant, pointing to a non-symmetrical and heavy-tailed distribution of residuals.

Overall Assessment:

- The model fits the GDP data reasonably well, but the non-normality and heteroskedasticity in the residuals suggest potential limitations in its accuracy.
- The significant negative coefficient of diff_log_unemploy highlights the interconnectedness between unemployment and GDP growth.
- The weaker autoregressive effects and less pronounced moving average terms compared to the unemployment model suggest a potentially more complex dynamic in GDP fluctuations.



As mentioned earlier, we have visualized the Multivariate ARIMA (AutoRegressive Integrated Moving Average) model by plotting it alongside the actual observed values. In the process of evaluating the model's performance, we calculated the Root Mean Squared Error (RMSE) as a metric to quantify the accuracy of the model's predictions. The calculated RMSE error for the predicted values was found to be 0.1103.

The RMSE is a statistical measure that gauges the average magnitude of the differences between predicted and observed values. In the context of time series forecasting, a lower RMSE value signifies a better fit of the model to the actual data. In this case, the obtained RMSE of 0.1103 suggests that, on average, the Multivariate ARIMA model's predictions deviate by approximately 0.1103 units from the actual observed values.

This information provides a quantitative assessment of the model's predictive accuracy, aiding in the understanding of its performance and the degree of precision achieved in forecasting the multivariate time series data under consideration. Further analysis and refinement of the model may be pursued based on these findings to enhance its forecasting capabilities.

e. Summary Evaluation

The presented output displays the Root Mean Squared Error (RMSE) values for three different forecasting models—VAR (Vector Autoregression), ARIMA (AutoRegressive Integrated Moving Average), and Multivariate ARIMA—applied to forecast Unemployment. Each RMSE value serves as a metric to assess the accuracy of the respective forecasting model in predicting Unemployment, with lower values indicating better performance.

	RMSE
VAR	0.100112
ARIMA	0.096952
Multivariate ARIMA	0.110335

- **VAR Model (RMSE: 0.100112):**

The VAR model leverages the relationships between multiple variables to make predictions. In this context, the RMSE of 0.100112 suggests that, on average, the VAR model's Unemployment predictions deviate by approximately 0.100112 units from the actual observed values. The lower the RMSE, the better the VAR model aligns with the true Unemployment data.

- **ARIMA Model (RMSE: 0.096952):**

The ARIMA model, which focuses on time series analysis and trends, yields an RMSE of 0.096952. This indicates that, on average, the ARIMA model's Unemployment predictions exhibit a deviation of approximately 0.096952 units from the observed values. The smaller RMSE value suggests a relatively accurate forecasting performance.

- **Multivariate ARIMA Model (RMSE: 0.110335):**

The Multivariate ARIMA model, which incorporates multiple time series variables, produces an RMSE of 0.110335. This value implies an average deviation of approximately 0.110335 units between the Multivariate ARIMA model's Unemployment predictions and the actual values. Though slightly higher than the other models, this RMSE still provides a measure of accuracy for the forecasting performance.

In summary, the RMSE values allow for a quantitative comparison of the forecasting accuracy among the VAR, ARIMA, and Multivariate ARIMA models. Analysts and decision-makers can use these metrics to assess which model best captures the underlying patterns in the Unemployment time series data, guiding further model refinement or selection based on performance metrics.

5. Conclusion and Recommendation

a. Conclusion

- Study aims to employ multivariate analysis to forecast the US unemployment rate and compare three prominent models: Vector Autoregressive (VAR), ARIMA-GARCH, and ARIMA. The study delves into the significance of accurate unemployment forecasting for economic policy making and business planning, providing insights into the interconnected dynamics of the labor market and the broader economy. The dataset includes variables such as the US unemployment rate, GDP, and inflation.
- The study starts with an introduction to the importance of unemployment forecasting and the need for multivariate models due to the complexity of economic systems. It outlines the business context, objectives, and the dataset's features. The Data Preparation and Exploratory Data Analysis section focuses on the essential libraries used and the handling of missing values and duplicates. The study then uses ACF & PACF plots to identify the stationarity of the unemployment rate and explores the relationship between unemployment and GDP through exploratory data analysis and correlation plots.
- The forecasting models include the Decomposition Method, Vector Autoregression (VAR), and ARIMA. The study evaluates the performance of each model, considering information criteria, parameter estimates, and diagnostic tests. It also highlights the Engle-Granger Test for Cointegration and ADF statistics for stationarity testing. Additionally, the study uses the RMSE metric to compare the accuracy of forecasting models, providing insights into the predictive performance of each model.

- In conclusion, the study provides valuable insights into the complexities of the US labor market and the relationships between unemployment and economic variables. It offers a comprehensive analysis of forecasting models, emphasizing their strengths and limitations. The study's detailed approach equips policymakers, businesses, and researchers with the tools necessary to understand and navigate the ever-changing US labor market. The study also recommends further research directions and provides a list of references for additional reading.
- Study serves as a valuable contribution to labor economics and forecasting, offering new insights into unemployment dynamics and equipping stakeholders with the necessary tools to make informed decisions.

b. Recommendation

The recommendation for the next study of this journal would be to explore potential extensions and future research directions in the field of labor economics and forecasting. This could involve incorporating additional variables such as technological advancements, globalization trends, or policy interventions to further refine forecasts and enhance our understanding of unemployment dynamics. Additionally, the study could delve deeper into the theoretical underpinnings and practical considerations of each forecasting model, discussing their strengths and weaknesses, data requirements, and interpretability of results. This comprehensive analysis would equip readers with a nuanced understanding of how multivariate analysis can be applied to improve unemployment rate forecasting and shed light on the intricate relationships within the US labor market. Furthermore, the study could explore alternative methods to capture the nuances and fluctuations in the data more effectively, allowing for a more accurate modeling of the observed fluctuations.

6. Reference

- Time Series Analysis: With Applications in R - Book by Jonathan Cryer and Kung-sik Chan
- Forecasting: Principles and Practice - Book by George Athanasopoulos and Rob J. Hyndman
- Applied Time Series Analysis - Course by The Pennsylvania State University
- Time Series Analysis – Course by Department of Applied Mathematics and Computer Science, Technical University of Denmark
- Time Series Analysis – Book by James D. Hamilton
- Time Series Models for Business and Economic Forecasting – Book by Philip Hans Frances, Dick van Dijk Anne Opshoor
- Analysis of Integrated and Cointegrated Time Serier with R- Book by Bernhard Pfaff
- A Review on Outlier/Anomoly Detection in Time Series Data – Paper by Ane Blazques, Angel Conde, Usue Mori, Jose A. Lozano
- Anomaly Detection: A Survey – Paper by Varun Chandola, Arindam Banarjee, Vipin Kumar
- Forecasting at Scale – Paper by Sean J. Taylor and Benjamin Letham