Time-Series Data Anaysis
: Forecasting the US Unemployment Rate

Group 13

Minjeong Song (2024710117)
Donghyun Jeon (2024711725)

June 3, 2024

Contents

About Data

- Unrate : Unemployment Rate in the US (1955.05 - 2024.04)
- UE : Employment Level - Part-Time for Economic Reasons in the US (1955.05 - 2024.04)
- First, split data into train and test subsets
  - train set : 1955.05 - 2017.05
  - test set : 2017.06 - 2024.04

Model Identification

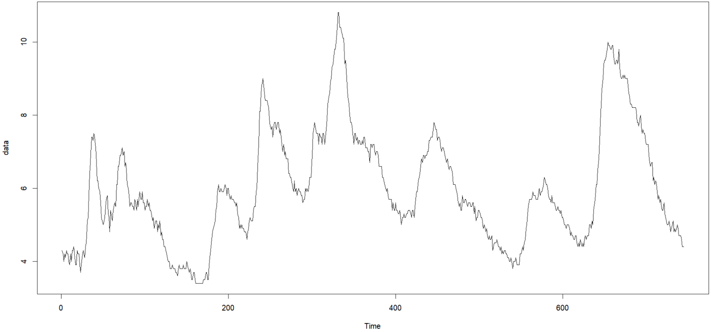


Figure: Time-series plot of Unrate

- Variance is unstable, and trend exists

Model Identification

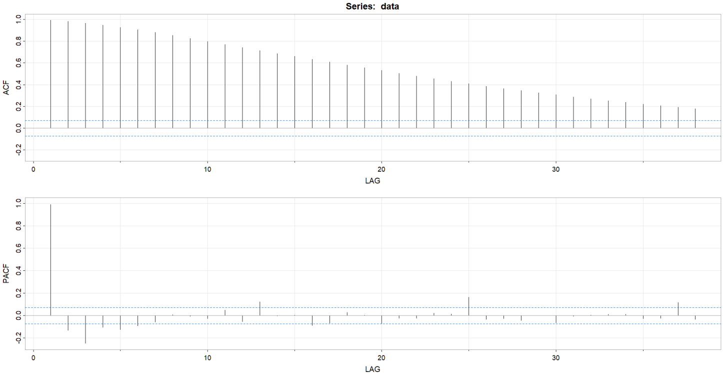


Figure: Correlograms of Unrate

- ACF of Unrate is slow-decaying

Model Identification

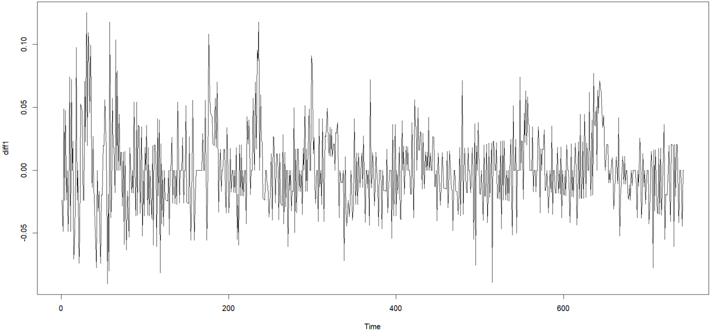


Figure: Time-series plot of $\nabla \ln y_t$

- Trend is eliminated

Model Identification

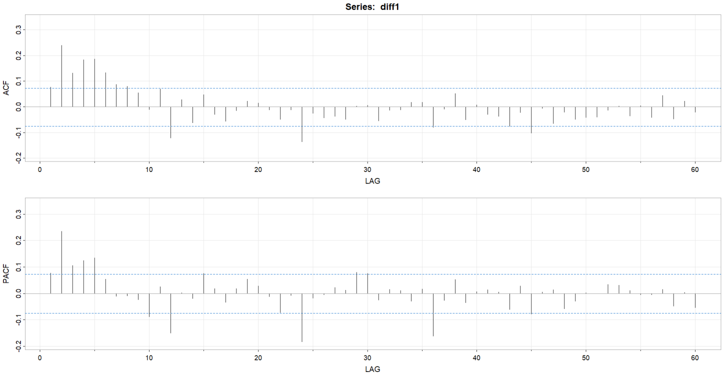


Figure: Correlograms of $\nabla \ln y_t$

Model Identification

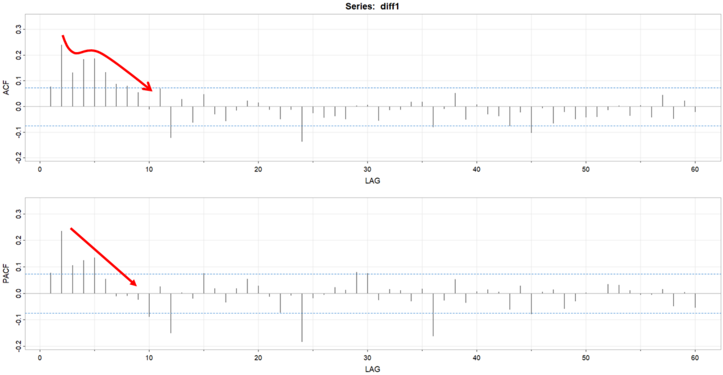


Figure: Correlograms of $\triangledown \ln y_t$

- ACF and PACF tails off (geometric decay) after lag 1

Model Identification

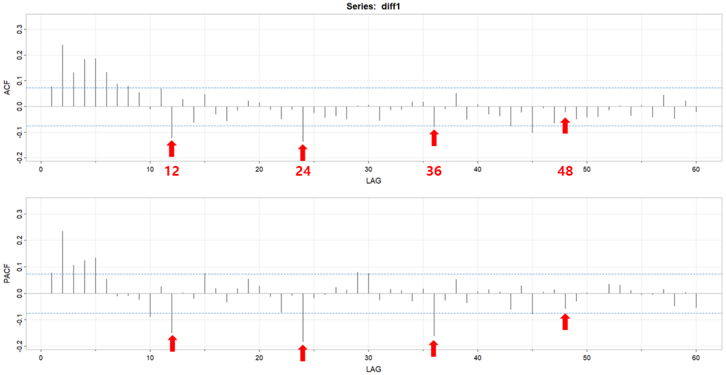


Figure: Correlograms of $\nabla \ln y_t$

- ACF and PACF tails off (geometric decay) after lag 12 or 24

Model Identification

- Before lag 12,

  - ACF : tails off (geometric decay) after lag 1

  - PACF : tails off (geometric decay) after lag 1

- Every 12 cycle,

  - ACF : tails off (geometric decay) after lag 12

  - PACF : tails off (geometric decay) after lag 12 or 24

- Consider Models

  - **SARIMA(1,1,1)(1,0,1)[12] or SARIMA(1,1,1)(2,0,1)[12]**

Model Identification

```
> sarima111101
Series: data_log
ARIMA(1,1,1)(1,0,1)[12]

Coefficients:
         ar1      ma1    sar1     sma1
      0.9306  -0.8142  0.5198  -0.8019
s.e.  0.0237   0.0341  0.0709   0.0507

sigma^2 = 0.0008621:  log likelihood = 1569.63
AIC=-3129.26   AICc=-3129.18   BIC=-3106.2
```

Figure: SARIMA(1,1,1)(1,0,1)[12]

```
> sarima111201
Series: data_log
ARIMA(1,1,1)(2,0,1)[12]

Coefficients:
         ar1      ma1    sar1     sar2     sma1
      0.9323  -0.8159  0.4706  -0.0602  -0.7377
s.e.  0.0236   0.0343  0.0995   0.0551   0.0939

sigma^2 = 0.0008619:  log likelihood = 1570.24
AIC=-3128.48   AICc=-3128.37   BIC=-3100.81
```

Figure: SARIMA(1,1,1)(2,0,1)[12]

Best SARIMA

Model	AIC	MSPE
ARIMA(1,1,4) by auto.arima	-3089.26	0.1063754
SARIMAX(1,1,1)(1,0,1)[12]	-3134.812	0.1050898
SARIMA(1,1,1)(2,0,1)[12]	-3128.48	0.1126912

- **Selected Model : SARIMAX(1,1,1)(1,0,1)[12]**
  $(1-0.9306B)(1-0.5198B^{12})(1-B)\ln y_t = (1-0.8142B)(1-0.8019B^{12})e_t$

How about SARIMAX?

- Add another data for exogenous variable
  - Employment Level - Part-Time for Economic Reasons (1955.05 - 2024.04)
  - Identically split data into train and test subsets
    - train set : 1955.05 - 2017.05
    - test set : 2017.06 - 2024.04

Model Identification

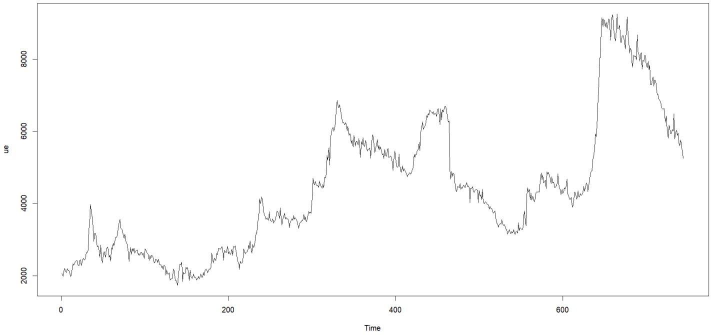


Figure: Time-series plot of UE

- Variance is unstable, and trend exists

SARIMAX

```
> sarimax111101
Series: data_log
Regression with ARIMA(1,1,1)(1,0,1)[12] errors

Coefficients:
         ar1      ma1    sar1     sma1    xreg
      0.9321  -0.8274  0.5262  -0.8085  0.0622
s.e.  0.0240   0.0341  0.0688   0.0488  0.0227

sigma^2 = 0.0008544:  log likelihood = 1573.41
AIC=-3134.81   AICc=-3134.7   BIC=-3107.14
```

Figure: SARIMAX(1,1,1)(1,0,1)[12]

Model Selection

Model	AIC	MSPE
SARIMA(1,1,1)(1,0,1)[12]	-3129.261	0.1150229
SARIMAX(1,1,1)(1,0,1)[12]	-3134.812	0.1050898

- **Selected Model : SARIMAX(1,1,1)(1,0,1)[12]**

$$(1-0.9321B)(1-0.5262B^{12})(1-B)\ln y_t = (1-0.8274B)(1-0.8085B^{12})e_t + 0.0622x$$

Model Diagnosis

- Residuals Test
  - H_0 : Residuals are independently distributed

```
> test(sarimax111101$residuals)
Null hypothesis: Residuals are iid noise.
Test                          Distribution Statistic   p-value
Ljung-Box Q                  Q ~ chisq(20)     43.17    0.0019 *
McLeod-Li Q                  Q ~ chisq(20)    123.28         0 *
Turning points T (T-495.3)/11.5 ~ N(0,1)       514    0.1044
Diff signs S          (S-372)/7.9 ~ N(0,1)       363    0.2537
Rank P         (P-138570)/3392.5 ~ N(0,1)    137753    0.8097
```

Figure: Residuals Test

  - H_0 is rejected
- This model is not appreciate for this data

Time-series Plot and CCF

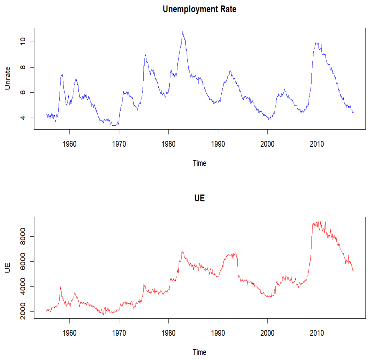


Figure: Time-Series of Unrate and UE

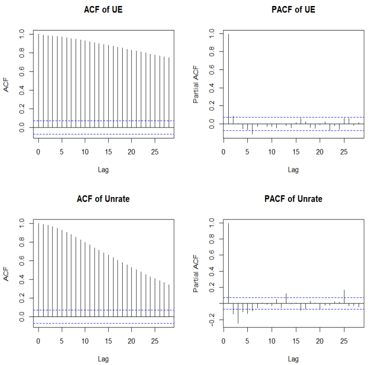


Figure: ACF of Unrate and UE

- Need log-transformation
- Gradually decreasing → need Lag-1 differencing

Prewhitening

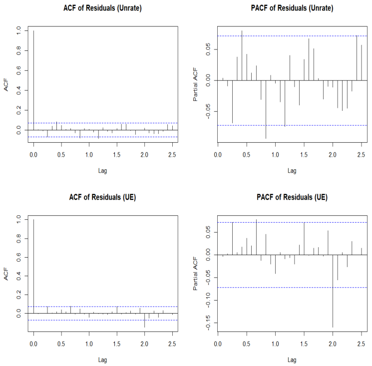


Figure: ACF and PACF of residuals of Unrate and UE

- Unrate $\rightarrow$ SARIMA(1,1,2)(2,0,1)[12]
- UE $\rightarrow$ MA(1)

CCF

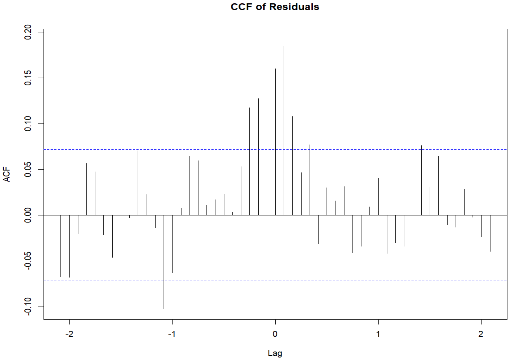


Figure: CCF between Unrate and UE

- VAR(2) or VAR(3)

VAR Model Comparison

Model	AIC	MSPE(UNRATE)	MSPE(UE)
VAR(2)	-5671.005	0.02116348	0.00985103
VAR(3)	-5673.634	0.02116387	0.00984801
VAR(4)	-5690.278	0.02116537	0.00984607
VAR(5)	-5700.318	0.02116474	0.00984560
VAR(6)	-5696.679	0.02116044	0.00984607

- We expect p=2 or p=3 for best parameter
- Select parameter through "VARselect" function in R
- AIC : 6, HQ : 5, SC : 4, FPE : 6

VAR(5) Model Notation

- A VAR(5) model is a model in which two time-series data are described using up to five lags. The formula for each time-series data is as follows::

$$\begin{aligned}\text{diff1}_t = &-0.0696\text{diff1}_{t-1} + 0.1536\text{diff2}_{t-1} + 0.0844\text{diff1}_{t-2} + 0.1277\text{diff2}_{t-2}\\ &- 0.0021\text{diff1}_{t-3} + 0.0852\text{diff2}_{t-3} + 0.0782\text{diff1}_{t-4} + 0.1197\text{diff2}_{t-4}\\ &- 0.1130\text{diff1}_{t-5} + 0.0113\text{diff2}_{t-5} + 0.0006\end{aligned}$$

$$\begin{aligned}\text{diff2}_t = &0.4122\text{diff1}_{t-1} - 0.2958\text{diff2}_{t-1} + 0.2706\text{diff1}_{t-2} - 0.1842\text{diff2}_{t-2}\\ &0.1806\text{diff1}_{t-3} - 0.0873\text{diff2}_{t-3} + 0.1014\text{diff1}_{t-4} - 0.1014\text{diff2}_{t-4}\\ &- 0.0021\text{diff1}_{t-5} + 0.0844\text{diff2}_{t-5} + 0.0015\end{aligned}$$

Model Diagnosis

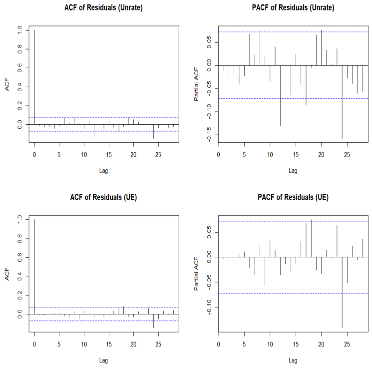


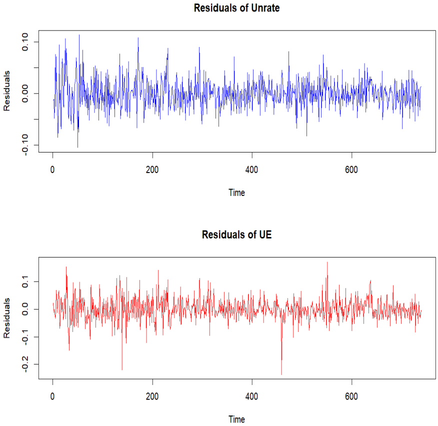
Figure: Residuals of ACF and PACF

```
        Box-Ljung test

data:  resi[, 1]
X-squared = 25.524, df = 12, p-value =
0.01253

> Box.test(resi[, 2], lag = 12, type = "Ljung-B
ox")

        Box-Ljung test

data:  resi[, 2]
X-squared = 6.1566, df = 12, p-value =
0.908
```

Figure: Ljung-Box Test

- Very little autocorrelation
- Residuals of UE are i.i.d.

Model Diagnosis



- ARCH test result $\rightarrow$ ARCH effect exists
- Heteroskedasticity

Introduction to Prophet

Prophet:

- Developed by Facebook as an open-source forecasting tool
- Handles seasonal changes and trends effectively with minimal data preprocessing
- Considers annual, weekly, daily patterns, and holiday effects for future predictions
- Easy to use and simple to model, making it a suitable choice for our application

Prophet

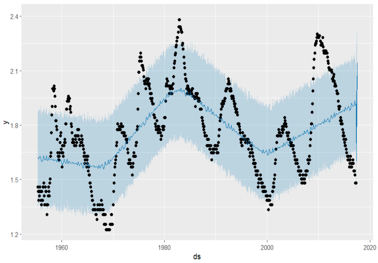


Figure: train (1955.05 - 2017.05)

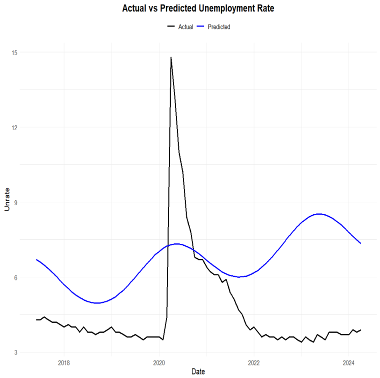
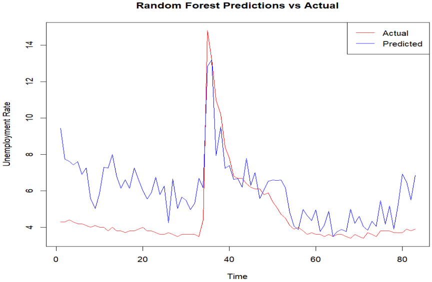


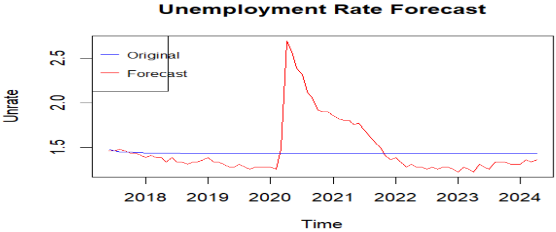
Figure: prediction (2017.06 - 2024.04)

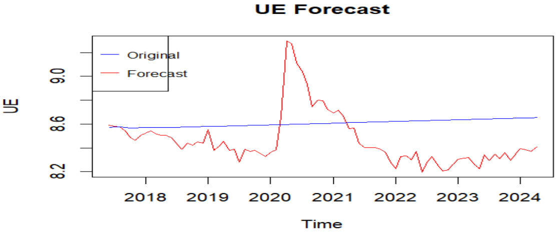
Random Tree Model



- Random Forest is proposed by Leo Breiman and Adele Cutler
- This model predicts very well around covid-19

Forecasting





- Unrate and UE were high in 2020 due to COVID-19

Final Model Selection

Model	AIC	MSPE(UNRATE)
VAR(5)	-5700.318	0.02116474
SARIMAX(1,1,1)(1,0,1)[12]	-3134.812	0.1050898

- We select VAR(5) for final model
- Unfortunately, because heteroscedasticity exists, it is necessary to consider the ARCH/GARCH model for better analysis

Final Model Selection

- We chose the VAR(5) model for its effectiveness in capturing the dynamic relationships among the variables
- However, the model did not adequately address the issue of heteroscedasticity, which was disappointing
- Additionally, the forecasting performance of the VAR(5) model was less impressive compared to machine learning methods

Thank you!