UNITED STATES UNEMPLOYMENT RATE FORECASTING

ECON – 5337 – 002
BUSINESS & ECON FORECASTING

INTRODUCTION

According to the Federal Reserve of St Louis database, the unemployment rate is the number of unemployed as a percentage of the labor force. The labor force represents anyone 16 or older residing in one of the 50 states or the District of Columbia who does not live in institutions (correctional or mental health facilities) or is not on active duty in the Armes Forces. The unemployment rate is the proportion of the workforce that is unemployed. Forecasting the unemployment rate for any country is very important as it helps predict the direction of the economy. Given everything going on in the world right now, the global pandemic, and the war in Ukraine, we are interested in forecasting the unemployment rate in the United States for six months. We will test two different forecasting methods and determine which one fits our data best. Then we are going to forecast the unemployment rate from the period of December 2021 through May 2022. Then we will compare our forecasted values to the actual unemployment rate from December 2021 to March 2022. We are anticipating a decline in the unemployment rate mainly because people who have left the workforce to retire early or focus on personal projects are likely to return to work because of the uncertainties we are currently facing.

LITERATURE REVIEW

In their article titled “Automatic time series modeling and forecasting: A replication case study of forecasting real GDP, the unemployment rate and the impact of leading economic indicators,” John Guerard, Dimitrios Thomakos & Foteini Kyriazi test and report on time series modeling and forecasting using several US. Leading economic indicators (LEI) as an input to forecasting the US. real GDP and the unemployment rate. The authors are interested in creating more statistically significant results using recently developed time series modeling techniques and software. They have the US. Leading economic indicators (LEI) data. An LEI is a piece of

economic data that corresponds with a future movement or change in some phenomenon of interest. It can help predict and forecast future events and trends in business, markets, and the economy. The technique used was Hendry and Doornik automatic time series PC-Give (AutoMetrics) methodology (Guerard, Thomakos, and Kyriazi, 2020). The conclusion was that Oxmetrics and Autometrics systems substantially reduce the regression sum of squares. The same findings had been addressed before; however, these results were more statistically significant (Guerard, Thomakos, and Kyriazi, 2020). They have found the best results for the univariate and bivariate models in terms of root mean squared error relative to naïve.

In summary, they reported a statistically significant impact of the LEI and weekly unemployment claims time series on real GDP and the unemployment rate series. The MZTT variable relationships are confirmed, in-sample and post-publication. These results were important since the unemployment rate was at a 50-year low in the U. S. Increasing economic activity drives down unemployment. “The LEI time series and its underlying methodology have evolved through the research of Burns and Mitchell (1946), Moore (1961), and Zarnowitz (1992). The LEI time series, and one of its components, weekly employment claims, is statistically significant in forecasting the unemployment rate”(Guerard, Thomakos, and Kyriazi, 2020). The authors have recreated the previous results more significantly, adding more value to the previous experiment. However, they could have used other methods to add more value and strength to the results.

In her article titled “Forecasting the unemployment rate using the degree of agreement in consumer unemployment expectations,” Claveria, Oscar aims to clarify unemployment forecasts by incorporating the degree of consensus in consumers’ expectations. She wanted to improve forecast accuracy in most countries. The results reveal that the degree of agreement in

consumers' expectations contains useful information to predict unemployment rates, especially for the detection of turning points (Oscar 2019). The data used is from European Commission - Business and consumer surveys. Hyndman and Khandakar's ARIMA model for stepwise regression models with an algorithm in two steps is the methodology used in the paper. The results showed that the proposed indicator leads to an improvement in forecast accuracy in most countries. The indicator of disagreement also helps refine predictions. Thus, the level of agreement in consumer unemployment expectations contains useful information to forecast unemployment rates, especially for predicting the turning points detected by agents in advance. The variation of improvements across countries reflects differences in the explanatory power of the agreement indicators used as predictors; it can also show the differences in other country-specific factors that represent the heterogeneity in the respective labor markets and the predictive capacity of consumers (Oscar 2019). Extending the analysis to control for some of these factors is an issue left for further research.

The third article reviewed was an article by Dingdong Yi, Shaoyang Ning, Chia-Jung Chang & S. C. Kou titled "Forecasting Unemployment Using Internet Search Data via PRISM." The authors wanted to use internet users' Google search data to forecast US unemployment initial claims. Also, to compare the new forecasting method, PRISM, to other forecasting methods. Internet users' Google data consist of internet and online platforms generated data. According to the authors, "internet users' online search of unemployment-related query terms provides highly informative and real-time" (Yi, Ning, Chang, & Kou. 2021). They estimated that since the weekly US unemployment claims reported by the Department of Labor each week are delayed by one week, using internet users' online search data can be highly informative. Also, using PRISM to forecast the unemployment rate might provide more significant results than

conventional methods. The authors used weekly initial claims from the ST. Louis Federal Reserve bank database, FRED, and the real-time internet search data from Google Trends. The authors compared PRISM, RMSE, and MAE of four models: the Bayesian structural time series (BSTS), exponential smoothing with Box-Cox transformation, ARMA, trend, and seasonal components (BATS), exponential smoothing with Box-Cox transformation, ARMA, trend, and seasonal components with trigonometric representation (TBATS), and naïve method. The outcome of the research was that PRISM outperformed all the other methods the entire time. Real-time Google Trends data help forecast with PRISM and BSTS. However, the contribution of real-time data becomes less significant in forecasting future weeks (Yi, Ning, Chang, & Kou. 2021).

In our last article review, Francesco D'Amuri and Juri Marcucci sought to determine whether US monthly unemployment rate predictions can be improved using the Google index (GI), a leading indicator based on internet job-related searches performed through Google in their paper titled " The predictive power of Google searches in forecasting US unemployment" (D'Amuri and Marcucci, 2017). The authors believe that providing accurate predictors for the labor marking is crucial, given how important they are for both investors and policymakers. Thus, having predictors that perform better than the ones already used can be critical. They used the seasonally adjusted monthly unemployment rate from the Bureau of Labor Statistics (BLS), the weekly seasonally-adjusted IC released by the US Department of Labor. Employment expectations for the manufacturing and non-manufacturing sectors from the Institute for Supply Management's (ISM) Report on Business (EEMt and EENMt respectively), the current and six-month-ahead consumer expectations from the US Consumer Confidence survey of the Conference Board (CEt and CE6Mt respectively), and the index of economic policy uncertainty

proposed by Baker et al. (2016) (D'Amuri and Marcucci, 2017). D'Amuri and Marcucci compared multi-step pseudo out-of-sample forecasting performances of a variety of linear models and univariate autoregression (AR(p)) using root mean squared forecast error (RMSFE). The authors found convincing evidence that using the Google index to predict the unemployment rate is useful. Google-based models outperformed other predictors and performed particularly well around the turning point at the beginning of the Great Recession, with their relative performance stabilizing thereafter (D'Amuri and Marcucci, 2017).

DATA COLLECTION

Our dataset comes from kaggle.com, an online community of data scientists and machine learning practitioners. The data contains the seasonally unadjusted monthly unemployment rate from January 1948 to November 2021. The data also includes total unemployment, men's and women's unemployment rate, and information on subsets of the population like ages ranging from 16 to 55. We will use seasonally unadjusted unemployment rate data from the St Louis Federal Reserve database, FRED, as the control for our analysis.

- **Descriptive statistics**

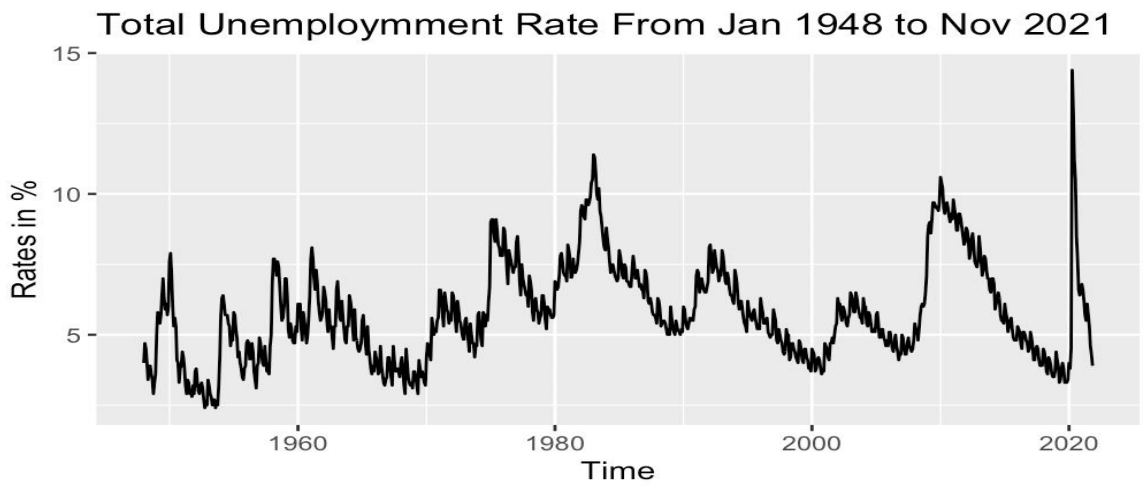
Variables	Description
Unrate	Total unemployment rate
unrate_men	Unemployment rate for men
unrate_women	Unemployment rate for women
unrate_16_17	unemployment rate for people aged 16 to 17
unrate_18_19	unemployment rate for people aged 18 to 19
unrate_20_24	unemployment rate for people aged 20 to 24

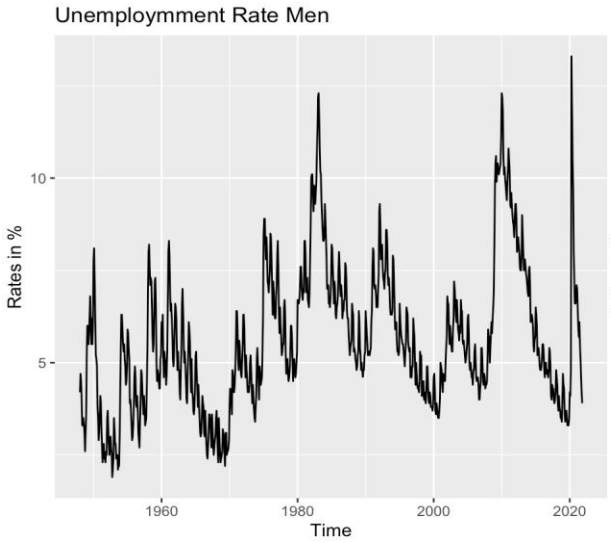
unrate_25_34	unemployment rate for people aged 25 to 34
unrate_35_44	unemployment rate for people aged 35 to 44
unrate_45_53	unemployment rate for people aged 45 to 54
unrate_55+	unemployment rate for people aged 55 plus

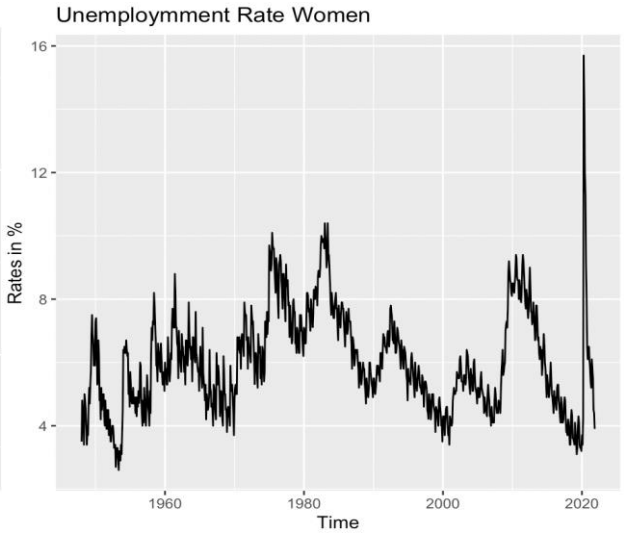
METHODOLOGY

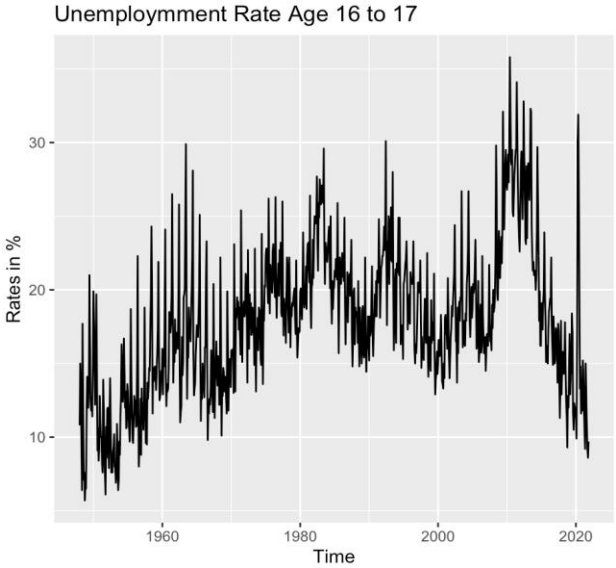
The unemployment rate is highly influenced by the business cycle and the overall health of the economy. As we have recently noticed during the pandemic, the unemployment rate skyrocketed as businesses were forced to close. We are interested in using different forecasting methods to forecast the unemployment rate. The first forecasting method we are using is time series regression. we will be regressing ($y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t$) the total unemployment rate as a function of the unemployment rate of the subsets groups in our data which are: men, women, and age groups 16 to 17, 17 to 18, 19 to 24, 25 to 34, 35 to 44, 45 to 55, and 55 and older. The second forecasting method we will examine is the AutoRegressive Integrated Moving Average (ARIMA (p,d,q)) which is a model that makes the dependent variate (y) a function of the last p lags of y, its last q lagged errors and d degrees of first difference.

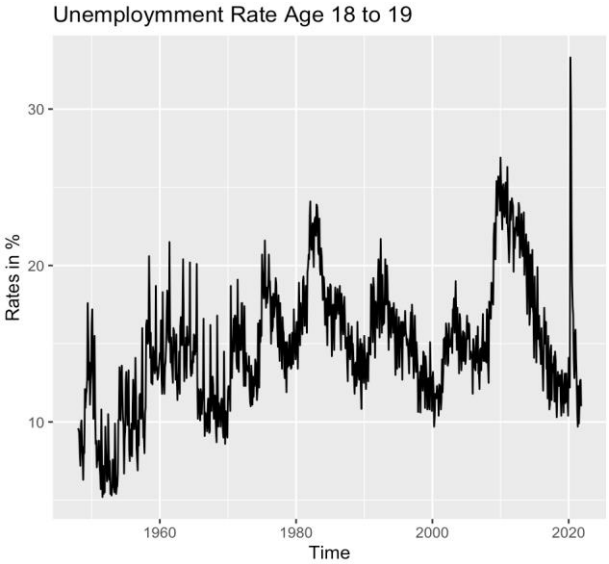
After plotting the data, we do not observe any particular instability in the variance. There is a slightly noticeable upward trend and seasonality. The most obvious feature of the data is that it is very cyclical. This is predictable because unemployment is sensitive to economic cycles as we have mentioned before. Because there is no variation of the variance, there is no transformation needed. After testing for unit root and seasonal variance differences, no seasonal difference nor first difference is necessary to obtain stationary data.

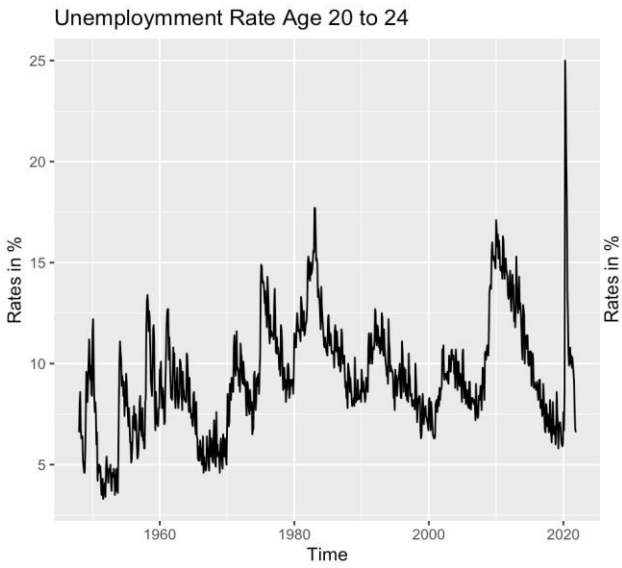

Total Unemploymment Rate From Jan 1948 to Nov 2021
Rates in %
15
10
5
1960
1980
2000
2020
Time

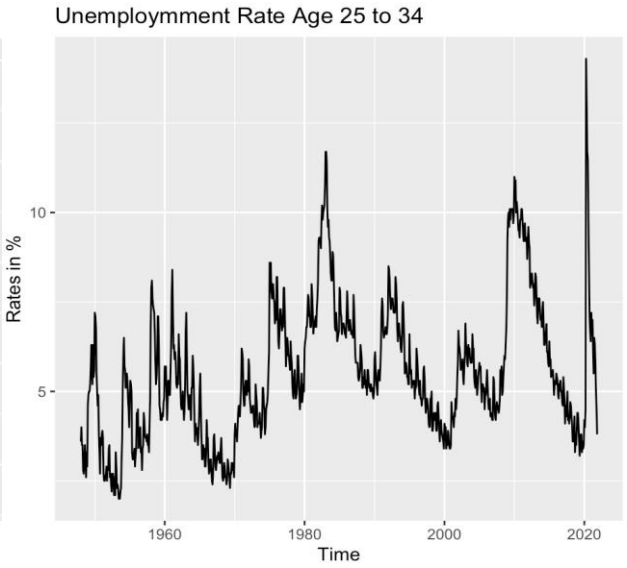

Unemploymment Rate Men
Rates in %
10
5
1960
1980
2000
2020
Time

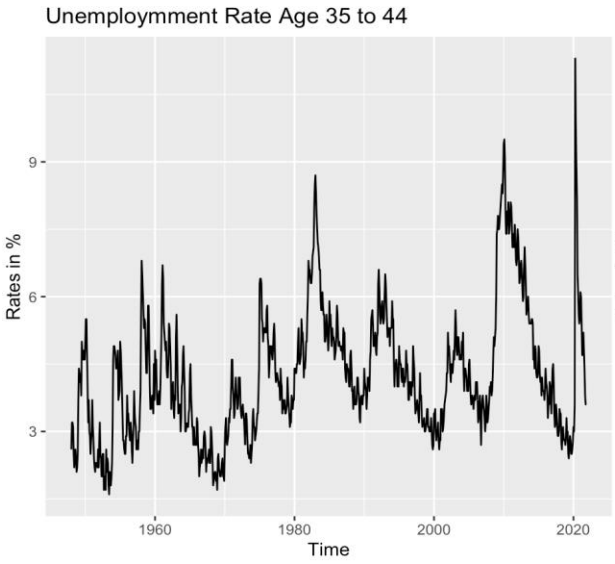

Unemploymment Rate Women
Rates in %
16
12
8
4
1960
1980
2000
2020
Time

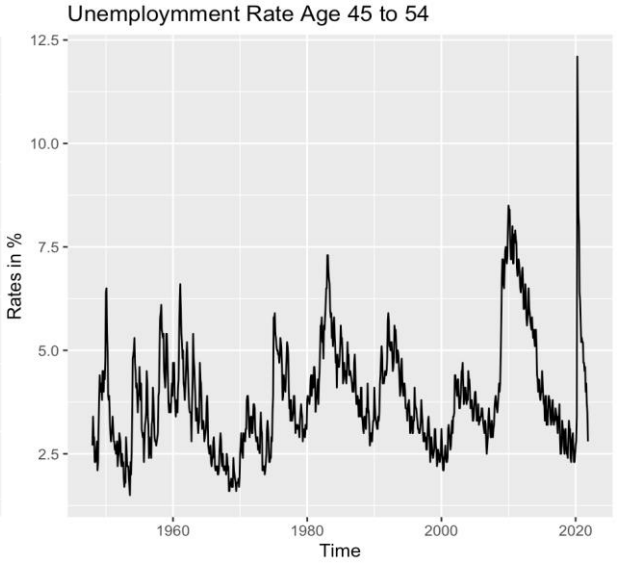

Unemploymment Rate Age 16 to 17
Rates in %
30
20
10
1960
1980
2000
2020
Time

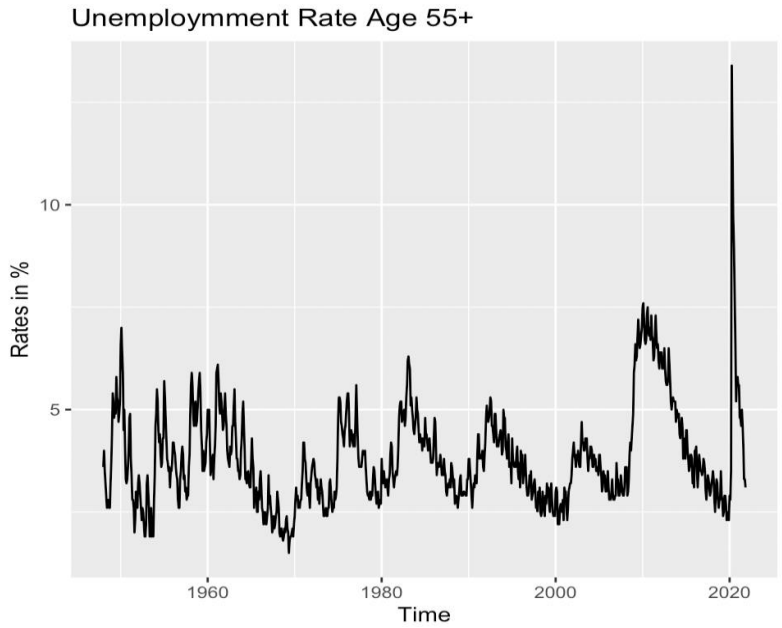

Unemploymment Rate Age 18 to 19
Rates in %
30
20
10
1960
1980
2000
2020
Time


Unemployment Rate Age 20 to 24
Rates in %
25
20
15
10
5
1960
1980
2000
2020
Time


Unemployment Rate Age 25 to 34
Rates in %
10
5
1960
1980
2000
2020
Time


Unemployment Rate Age 35 to 44
Rates in %
9
6
3
1960
1980
2000
2020
Time


Unemployment Rate Age 45 to 54
Rates in %
12.5
10.0
7.5
5.0
2.5
1960
1980
2000
2020
Time



- Summary Statistics

Variable	Obs	Mean	Std. dev.	Min	Max
unrate	887	5.763134	1.740101	2.4	14.4
unrate_men	887	5.633709	1.954639	1.9	13.3
unrate_women	887	6.028749	1.608252	2.6	15.7
unrate_16~17	887	17.94352	5.018894	5.7	35.8
unrate_18~19	887	14.8248	4.047867	5.2	33.3
unrate_20~24	887	9.34566	2.800988	3.3	25
unrate_25~34	887	5.532582	1.9236	2	14.3
unrate_35~44	887	4.242954	1.443626	1.6	11.3
unrate_45~54	887	3.867193	1.352247	1.5	12.1
unrate_55_~r	887	3.838782	1.241579	1.5	13.4

- **Time Series Regression**

Stepwise regression is the iterative construction of a regression model in which the independent variables to be used in the final model are chosen one by one. It entails incrementally adding or eliminating potential explanatory factors, with each iteration requiring statistical significance assessment. Similarly, we have taken individual variables first and

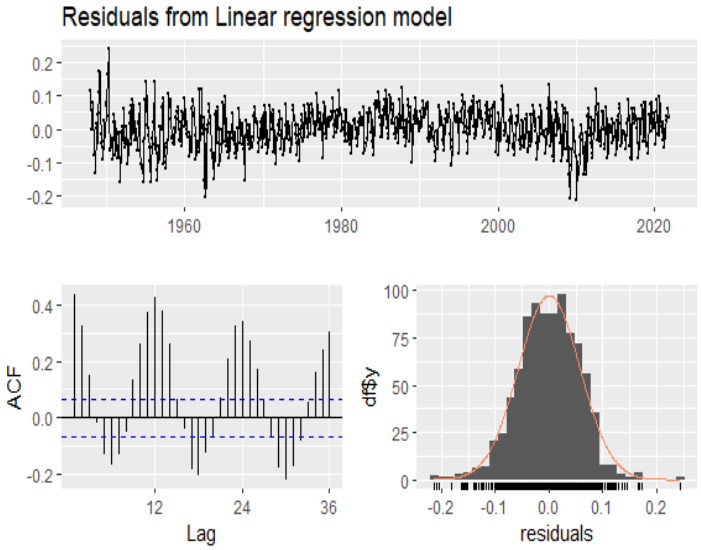
checked which variable is having lowest AICc values while training the model. And adding one-by-one variables to that model we come to know that the model which performs best based on AICc involves all the explanatory variables. We have used the following time-series equation to predict our unemployment rate.

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

Where y_t represents the total unemployment rate,

x_{1t}, $x_{2t,\ldots,}$ x_{kt} = men, women, and age groups from 16 to 55+

We ended up with the model where all the independent variables are included. And with help of the residual plot, we can say that the residuals are white noise for our selected model and it fits the assumptions. So, we can forecast further using the same model.



- **ARIMA**

The plotting the data, the ACF is exponential decay, and PACF plots have a significant spike in lags 10 and 12. The selected starting model is ARIMA(10,0,0)(1,0,0) with drift. After testing several models around, the best model we landed on was an ARIMA(8,0,1)(4,0,0) with an AICc (Corrected Akaike's Information Criterion) of 1081.69.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \phi_5 y_{t-5} + \phi_6 y_{t-6} + \phi_7 y_{t-7} + \phi_8 y_{t-8} + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \Phi_3 y_{t-3} + \Phi_4 y_{t-4} + \Theta \varepsilon_{t-1} + \varepsilon_t$$

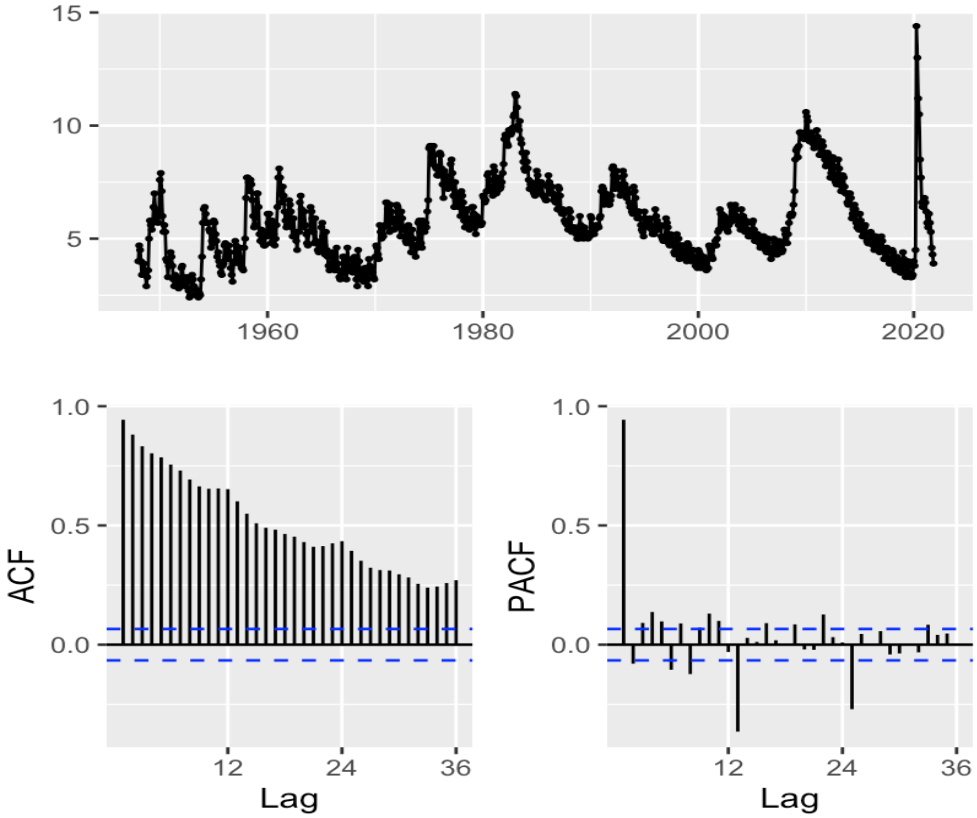
Where y_t = unrate (total unemployment rate) c = constant

Φ = autoregressive coefficient

Φ = seasonal autoregressive coefficient

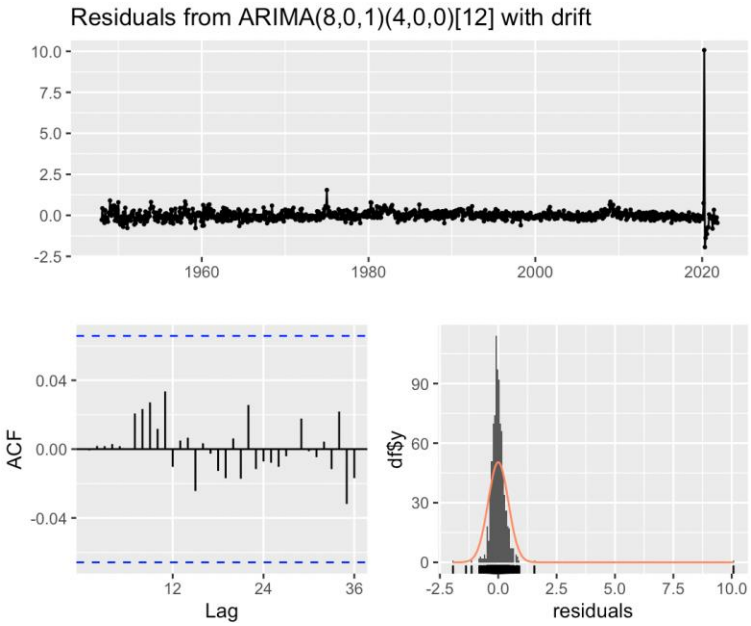
θ = seasonal moving average coefficient

ε = error term



The residuals are white noise and normally distributed. There is an outlier on the residual distribution plot but it might be because of the pick in unemployment in 2020. There is no evidence of serial correlation as the P-value of the Ljung-Box test is 0.84. This means there is no information contained in the residuals that could help improve our forecast

However, auto.arima picked an ARIMA(3,1,3)(2,0,0) with drift. Because we cannot compares those two models using AICc since one has a seasonal difference and the other does not. The residuals are normally distributed and white noise, there are two significant lags out of 36, which is fine since we are looking at the 95% interval. There is some evidence of serial correlation as the P-value of the Ljung-Box test is 0.048 in this model, slightly lower than the 0.05 needed to reject the hypothesis of serial correlation. Using Root Mean Squares Errors (RSME) measure. The model that fits the data best is the one we generated. The ARIMA(8,0,1)(4,0,0) with a RMSE of 0.4323785 compared to 0.4567895 for the model generated by auto.arima.



RESULTS

- **Stepwise Regression:**

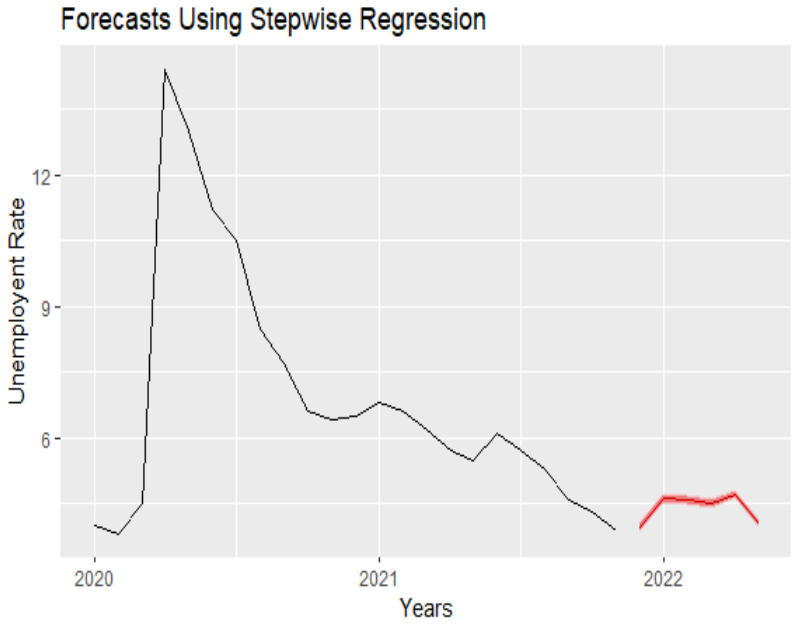
The equation we received after the regression with coefficients will be:

*unrate =-0.041 + 0.55*unrate_men + 0.3882423*unrate_women - 0.0002*unrate_16_to_17 - 0.003*unrate_18_to_19 + 0.029*unrate_20_to_24 + 0.05*unrate_25_to_34 - 0.005*unrate_35_to_44 - 0.035*unrate_45_to_54 + 0.009*unrate_55_over*

Then we checked the residuals of the model and found that the residuals look like white noise and there was no evidence of serial correlation. So, we considered the same model for our forecasting.

For forecasting, we used the ETS (Error, Trend, Seasonal) method is a method for forecasting univariate time series. The trend and seasonal components are the emphases of this ETS model. The ETS model's versatility is based on its ability to trend and seasonal components of many attributes. So, we initially forecasted individual independent variables and then combined them to forecast our target variable. And the results seem reasonable. The generated forecast is illustrated below.

```
Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Dec 2021       3.949921 3.875510 4.024332 3.836048 4.063794
Jan 2022       4.612700 4.538232 4.687168 4.498739 4.726661
Feb 2022       4.571602 4.497247 4.645956 4.457815 4.685388
Mar 2022       4.517366 4.443037 4.591695 4.403619 4.631114
Apr 2022       4.720859 4.646454 4.795265 4.606995 4.834724
May 2022       4.040660 3.966260 4.115061 3.926803 4.154517
```



Our forecast showed that the unemployment rate would increase for the six months after November 2021. The real unemployment rate values from the Fred database reveal that unemployment rose marginally in December 2021 before dropping from January to March 2022. While our forecasted values are similar for December 2021. However, it changes for 2022 as it is approximately steady and then declined after May 2022.

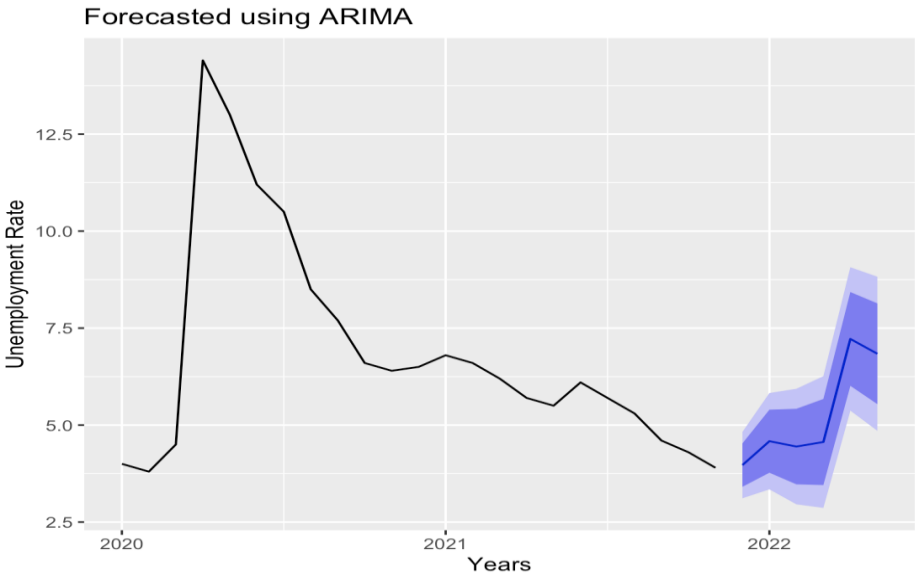
- ARIMA

After using the selected Arima model to forecast the unemployment rate for the next six months after November 2021. The estimated forecasting equation is:

$$y_t = 5.073 + 2.052y_{t-1} - 1.201y_{t-2} + 0.247y_{t-3} - 0.206y_{t-4} + 0.214y_{t-5} - 0.185y_{t6} + 0.109y_{t-7} - 0.032y_{t-8} + 0.074y_{t-1} + 0.294y_{t-2} + 0.245y_{t-3} + 0.261y_{t-4} - e_{t-1} + 0.002$$

The generated forecast are illustrated bellow.

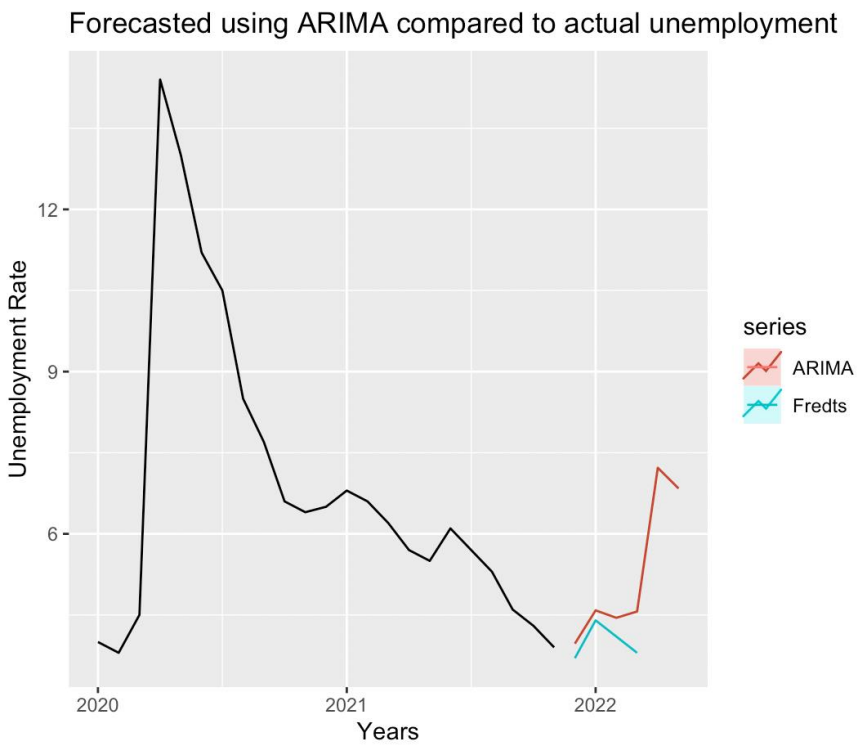
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Dec 2021	3.968961	3.409902	4.528021	3.113954	4.823969
Jan 2022	4.584662	3.772794	5.396530	3.343017	5.826308
Feb 2022	4.446665	3.473319	5.420012	2.958060	5.935271
Mar 2022	4.563881	3.454259	5.673502	2.866861	6.260900
Apr 2022	7.220564	6.011988	8.429141	5.372206	9.068922
May 2022	6.839231	5.540904	8.137558	4.853611	8.824851

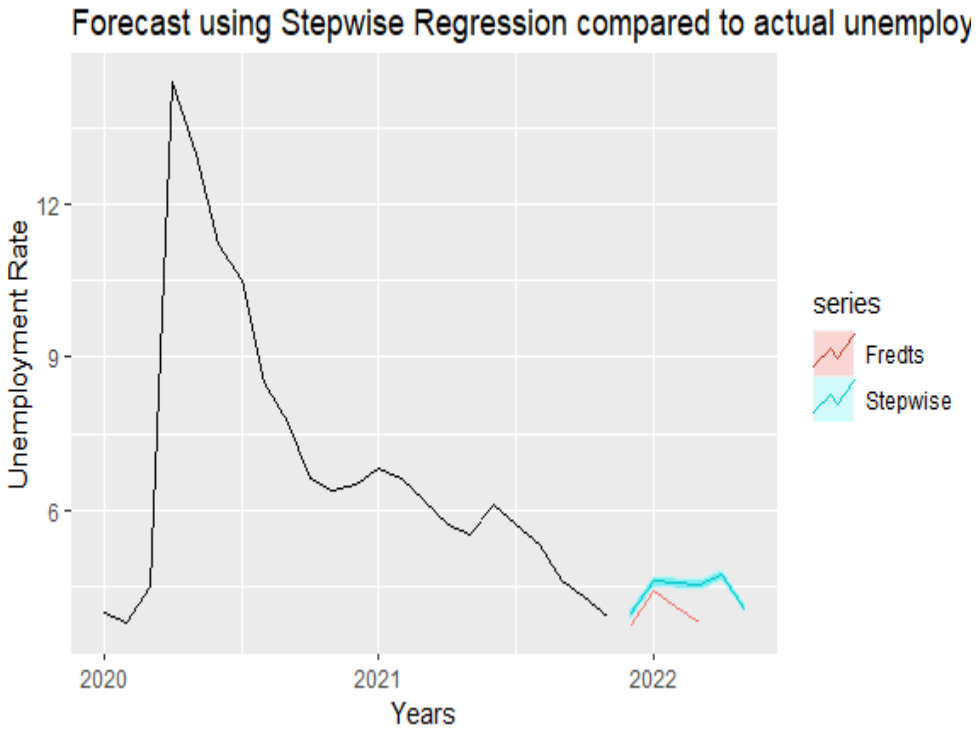


Furthermore, after examining the accuracy of all of the models, we discovered that stepwise regression had the highest accuracy of all of them. The accuracy results for stepwise regression and Arima were as follows:

```
> accuracy(Stepwise)
                      ME       RMSE        MAE         MPE      MAPE      MASE      ACF1
Training set 9.839715e-19 0.05734335 0.04555033 -0.02851233 0.8576403 0.0480315 0.4359499
> accuracy(Arima_model)
                      ME      RMSE       MAE        MPE     MAPE      MASE         ACF1
Training set -0.002161969 0.4323732 0.2082022 -0.5356174 3.731916 0.2195432 -0.00106688
```

So, we can consider that the stepwise regression was the best model out of all the three models that we have used for our analysis. Our forecast showed that the unemployment rate would increase for the six months after November 2021. However, the actual values of the unemployment rate front the Fred database show that unemployment rose slightly in December 2021 and then dropped from January through March 2022. The difference between the forecasted values and the actual unemployment rate might be caused by the changes in economic conditions.


Forecasted using ARIMA compared to actual unemployment
Unemployment Rate
12
9
6
2020
2021
2022
Years
series
ARIMA
Fredts


Forecast using Stepwise Regression compared to actual unemploy
Unemployment Rate
12
9
6
2020
2021
2022
Years
series
Fredts
Stepwise

CONCLUSION

The unemployment rate in the United States saw a sharp increase during the global pandemic. A lot of companies were not able to afford to keep their consumers because of lockdowns. Also, people are quitting their jobs to look for better-paying jobs, focus on their projects, or retire early. With high inflation rates and the Russian attack on Ukraine, we were interested in testing different forecasting models to predict the unemployment rate. After finding the best model, we used it to predict the unemployment rate for six months. Our expectation was that the unemployment rate would go down. Our forecast showed that unemployment would increase from December through May 2022. However, the actual value of the unemployment rate rose in December 2021 but dropped from January through March 2022. Given the current employment conditions, we can expect the unemployment rate to drop again between April and May 2022. It is important to mention that since the beginning of this semester, the market conditions and monetary policies have changed. There is less worry about the unemployment rate in the current market, but there are more concerns about the inflation rate and its impact on the economy.

REFERENCES

Claveria, Oscar. (2019) Forecasting the unemployment rate using the degree of agreement in consumer unemployment expectations. Journal for Labour Market Research **53,** 3. DOI:10.1186/12651.019.0253.4

Dingdong Yi, Shaoyang Ning, Chia-Jung Chang & S. C. Kou (2021) Forecasting Unemployment Using Internet Search Data via PRISM, Journal of the American Statistical Association, 116:536, 1662-1673, DOI: 10.1080/01621459.2021.1883436

Francesco D'Amuri, Juri Marcucci (2017), The predictive power of Google searches in forecasting US unemployment, International Journal of Forecasting, 33, 4, 801:816, ISSN 0169-2070

John Guerard, Dimitrios Thomakos & Foteini Kyriazi | Xibin Zhang (Reviewing editor) (2020) Automatic time series modeling and forecasting: A replication case study of forecasting real GDP, the unemployment rate and the impact of leading economic indicators, Cogent Economics & Finance, 8:1, DOI: 10.1080/23322039.2020.1759483

U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, February 26, 2022.