Sanghyun Cho

San Francisco Bay Area Unemployment Rate 1 Year Forecast

Introduction

This paper will forecast the next 12 months' unemployment rate in the San Francisco-Oakland-Hayward (San Francisco Bay Area) metropolitan statistical area (MSA) from October 2020 to October 2021 using the data from the Bureau of Labor Statistics (BLS). The initial expectation of the unemployment rate is to decrease in the next 12 months, which is based on the expectation of increased economic activities and economic rebound after the development of vaccination.

According to the Current Population Survey (CPS), the unemployment rate is the number unemployed as a percent of the civilian labor force. Unemployed people are all person who, were available to work and made effort to find employment in the past 4 weeks, had no employment. The civilian labor force includes all the people in the civilian noninstitutional population ages 16 or above who are classified as either employed or unemployed. The civilian labor force, however, excludes military personnel, federal or government employees, handicapped or discouraged workers, and agricultural workers.

San Francisco Bay area is chosen as the location of forecast because of its high population, 12th highest population among 383 other MSAs according to US Census Bureau (USCB, 2019). This characteristic makes the unemployment rate change of San Francisco MSA affect a significant number of people. Because ones' unemployment can be a great risk to their households finance, the unemployment rate has a great meaning to society. Therefore, the

unemployment rate is an important economic indicator, and a precise forecast of the unemployment rate in metropolitan areas with high populations, such as the San Francisco Bay Area, can be invaluable information for states and governments when making policies.

In order to construct the forecast model, I will first explain the characteristics of the San Francisco Bay area unemployment rate data. Since the construction of the model depends heavily on the original data, it is important to clarify the characteristics. Important characteristics include trend, cycle, seasonality, autocorrelation, heteroskedasticity, and serial correlation.

Data

Figure1.

Figure 1 is a monthly, seasonally unadjusted data of the SF Bay Area unemployment rate from 1990 January to 2020 September. This data is gathered through the Survey of Local Area Unemployment Statistics Program (LAU) and retrieved from BLS. Since the data is not

seasonally adjusted, the seasonality is visible from the figure. In each year, there are two visible spikes in the unemployment rate. These spikes are seasonality and they repeat every year during a certain period of time and it is only affected the time.

If seasonality is small spikes, the cycle is a big flow of up and down that is visible in a longer period of time. For example, in figure1, from 1990, the unemployment rate is constantly increasing roughly for 4 years, then until 2000, the unemployment rate starts to decrease again. Interestingly, the magnitude of the cycle is different every time. In between 2008 and 2010, during the economic recession, the unemployment rate rose higher compared to different periods, like the early 1990s and early 2000s.

On the largest scale, we can see the trend. Although not clearly visible, for the last 30 years of the period, the unemployment rate seems to be slightly increasing. One way to observe a trend is to compare the unemployment rates at the deepest contractions or at the highest expansions. In this figure, the deepest point at contraction has been slowly moving upward over time. Therefore, the unemployment data has a slightly upward sloping trend. In general, it usually takes a longer time to clearly observe a trend.

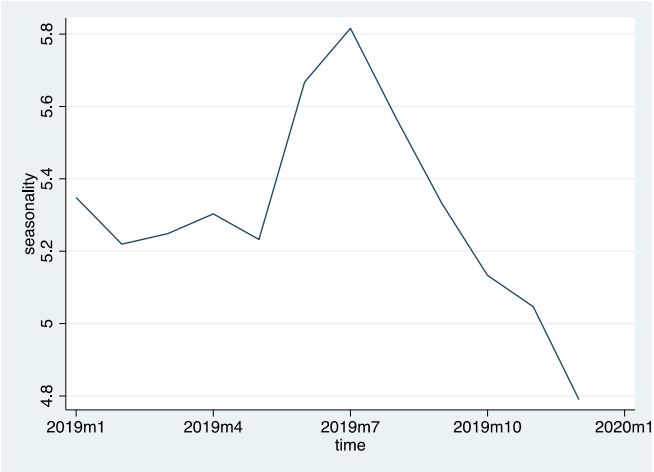


Figure 2.

Figure 2 is a seasonality model. By setting seasonal dummy variables, it captures the seasonality of the unemployment rate for every month in any year. So, the x-axis of the figure is a randomly selected year, and the seasonality will be consistent in whichever year is chosen for the x-axis. The big difference in seasonality at the end of the year and the beginning of the year shows that the unemployment rate will sharply increase during the winter. Possible explanations include the decreased construction during the winter and the increased lay-off during the long holidays in winter. Many employees working in the construction sector in northern states cannot work due to harsh weather and they have to look for different jobs during wintertime, such as snow removal.

Another important characteristic of this data is the autocorrelation. Autocorrelation means that the data is correlated with itself from the previous period. For example, in figure3, we can see that autocorrelation is starting from 1 at lag 0. As for lag increases, autocorrelation is decreasing and after roughly 40 lags, autocorrelation almost disappears. This is ergodicity, where autocorrelation approaches to zero over lag.

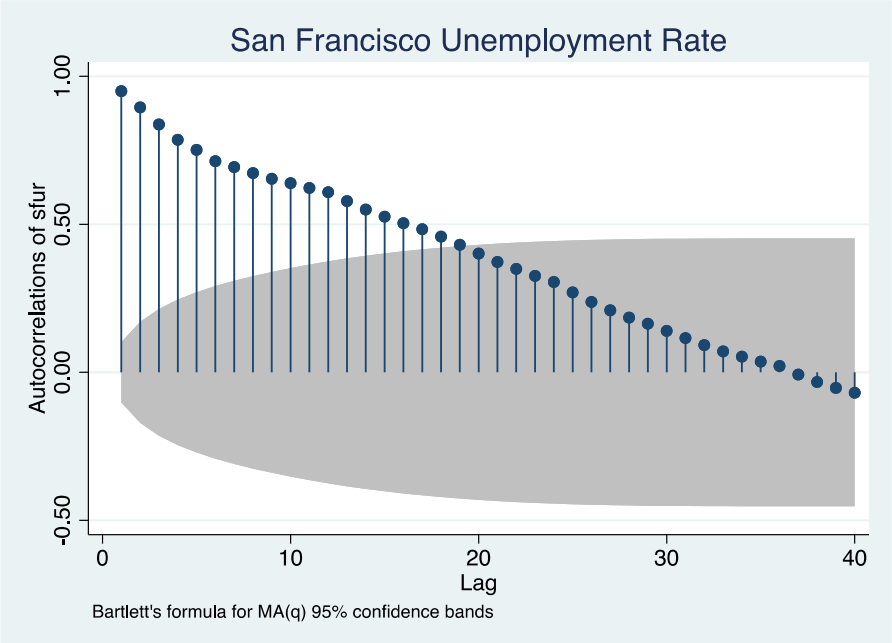


Figure 3.

On the other hand, autocorrelation is high for lags ranging from 0 to 20. This is a serial correlation, where the current value is correlated with the value of the previous time. It means that the current month's unemployment rate is strongly correlated with the previous month's unemployment rate. In this figure, if last month's unemployment rate was high, then this month's unemployment is likely to be high.

Another way to observe a serial correlation is to look at the distribution of residuals. Since the lagged value from the error term is affecting the variable, the residual is no longer randomly distributed. As seen in figure 4, serially correlated data will show a certain pattern. In this case, the residual increases and decreases over time and shows fluctuating patter in a long-term period.

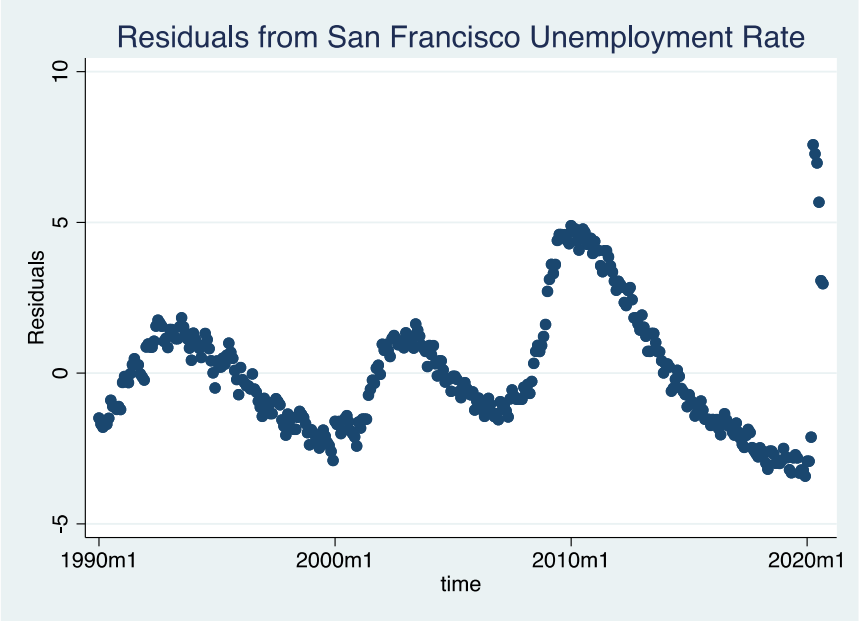
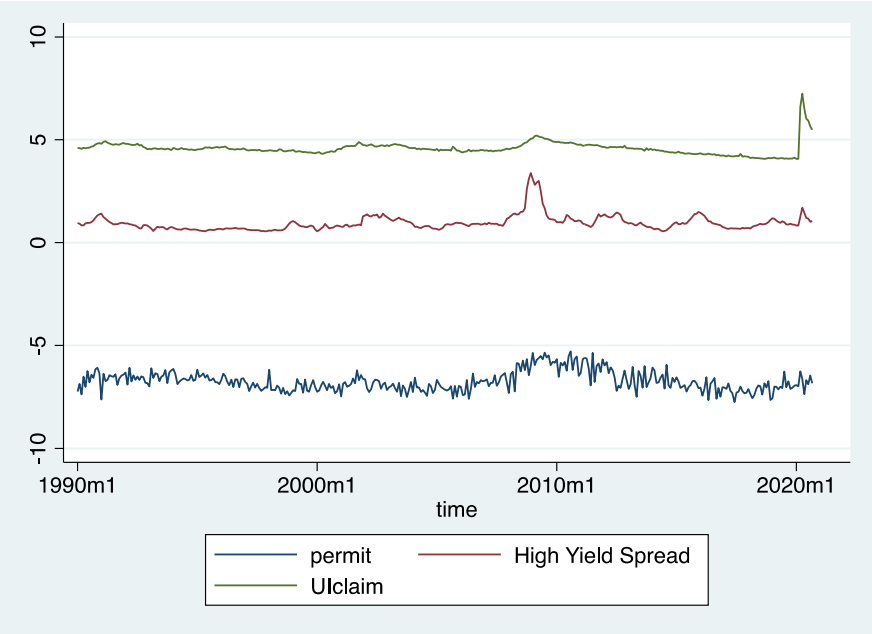


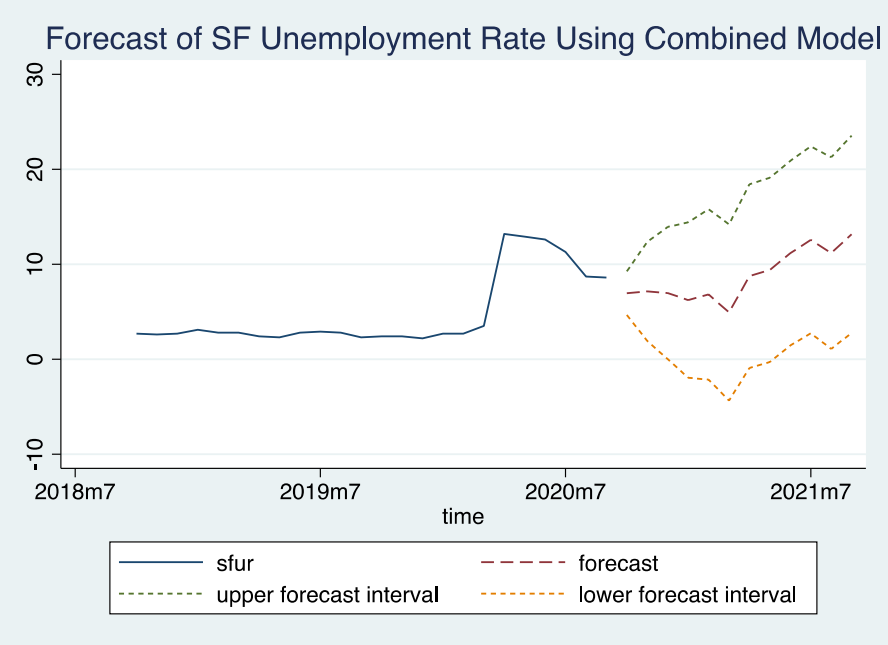
Figure4.

From figure4, we can also observe heteroskedasticity. A data is heteroskedastic when the variance of a variable is unequally distributed. In this figure, the fluctuation of residual is getting

steeper over time, and also the height of the peak is increasing. This shows that the variance of variables is changing over time. For heteroskedastic and serially correlated data, classical regression is not optimal. Therefore, when data is both heteroskedastic and serially correlated, we use the Newey-West estimator instead of classical data. However, if the data is only heteroskedastic, robust regression is the optimal choice.



Model1



Model2

Model 2 is forecasted with Autoregressive Distributed Lag (ADL) model. For the ADL model, I used 7 months lag autoregression on the San Francisco unemployment rate (sfur) with 3 different leading indicators including initial jobless claims for unemployment insurance by unemployed individuals, building permits, and corporation high yield spread. All the data of leading indicators are retrieved from the Federal Reserve Bank of St. Louis Data Base. Both unemployment insurance claims and building permits are seasonally adjusted data and taken into log format due to the large magnitude of change, particularly during the COVID-19 pandemic in 2020. All three data sets are formatted monthly from 1990 January to 2020 September.

Building permits are an adequate leading indicator because this data shows whether future construction will increase or decrease. There will be less construction during the recession, so decreased building permits reflect the increased possibility of a recession. Since unemployment increases during and after the recession, a building permit is a good leading indicator of the unemployment rate. Moreover, the construction sector takes a large proportion of employment. Therefore, a decreasing number of the construct will directly impact the unemployment rate to increase. For consistency with other leading indicators, all the values from the building permits dataset are inversed, so the large value from the original data set infers the increased possibility of economic recession, not growth.

Initial jobless claims by unemployed individuals are another good leading indicator. This data shows how many people, who want to work, are out of work at a given time. Therefore, it implies, in advance, how the unemployment rate will change. Generally, under the assumption that the economy is healthy when the job market is healthy, increased initial claims is a bad sign for the economy and therefore infers the higher unemployment rate. Despite the good characteristics as a leading indicator, the historically high level of claims recorded during the

COVID-19 pandemic, clearly visible in Model 1, makes this data complicated to use for the forecast of unemployment rate even after log adjustment. Therefore, this paper also provides an unemployment rate forecast omitting the initial jobless claims variable.

Corporation high yield bond spread is a difference between the rate of the low-grade bond, BAA, and high-grade bond, AAA. BAA rated corporates usually have a higher risk of bankruptcy compared to AAA-rated corporates. Therefore, if investors think there will be a recession, and think BAA firms have an increased chance of bankruptcy, investors will hold BAA bonds only when its interest rate increases. Although the BAA rate reflects the perceived risk of recession, the AAA rate also reflects the same risk. Therefore, the difference between BAA and AAA rate is a good leading indicator. An increase in the difference between BAA and AAA rate will reflect the perceived increased risk of the economic downturn which causes an unemployment rate to increase.

In addition to theoretic economic explanation, this paper also tests if these leading indicators are statistically adequate for forecasting unemployment. The p-value of the Granger prediction test for permits, initial claims, and high yield spread is 0.0089, 0, 0.0001 respectively. So with 95% confidence, I can reject the null hypothesis that it does not granger cause the unemployment rate. Therefore they are all statistically adequate leading indicators for this forecast.

To decide how many lags and variables to use, I use the Akaike Information Criterion (AIC). Since the main purpose of the test to find the optimal forecast model, AIC is a better criterion to use. In the case of AIC, which searches for an optimal forecast model, finding a good mean squared forecast error (MSFE) is more important than finding what true drives are. Therefore, in this case, using AIC is better than the Bayesian Information Criterion (BIC), which

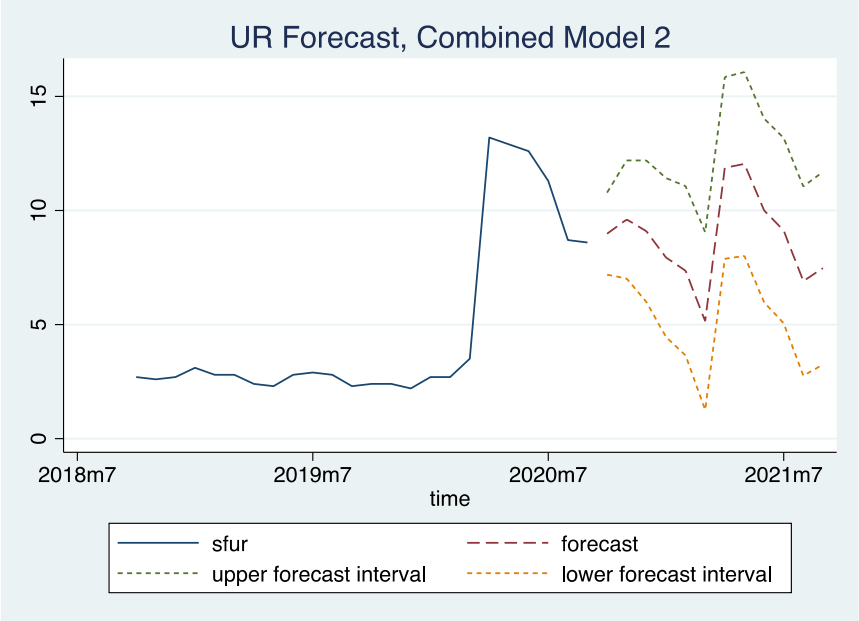
assumes that there is one true model. Using AIC to test for each of the leading indicators and autoregression of unemployment, the result suggests that for autoregression, it is optimal to use 8 months lag. For each leading indicators, I compare the AIC values to find out what is more suitable. Of all 3 indicators, initial claims with 12 months lag variable is the most suitable. However, under AIC, the model with combined indicators is an even better option for forecast modeling. Therefore, after multiple steps of AIC value comparison, model 2 is graphed with a combined model of 8months lag autoregressive unemployment rate, initial claims with 12 months lag, building permits with 6months lag, and high yield bond spread with 6 months lag.

Model 2 is the 1-year forecast of the San Francisco MSA unemployment rate. In this forecast, the expected unemployment rate is slowly decreasing from 7% to 5% until March of 2021. However, after March 2021, the expected unemployment rate starts to increase rapidly until the expected unemployment rate of 13% in September 2021. Although it is not easy to predict exactly when the unemployment rate will revert back to under 5 percent, considering that many firms are successfully testing the vaccines, the expected decrease of the unemployment rate until March 2021 seems plausible.

The expected sharp increase in the unemployment rate after March 2021, however, is questionable. The previous over 10 percent unemployment rate peak was when the COVID-19 began to spread and were uncontrollable. Therefore, it is not easy to expect the unemployment rate to increase back to 13% after march 2021. However, until the vaccine for COVID-19 is developed, there is always a chance of disease spreading and the unemployment rate fluctuating. The forecast interval increases from a lower bound of 4.6 percent and an upper bound of 9.2 percent in one month ahead of the forecast to 2.7 percent lower bound and 23 percent upper bound in the final month of the forecast. Although the forecast interval gets wider over time as

the variance and standard deviation increases, the interval shown in model 2 is very wide. One possible explanation for the wide interval in this model is the large number of variables used in this model. The model gets more precise and less biased as the number of variables increases, however, the variance will increase at the same time. This is clear when comparing

As a result, this model shows an expectation of a steeply increasing unemployment rate trend followed by a small decrease in the beginning. Although the model is precise, the forecast interval is very wide due to many variables, which makes this model very generous. In the early stage of the forecast, part of the lower bound of the forecast is under 0. This is not realistic because the level of unemployment cannot be a negative number. To address these issues, I also include another forecast model with less variable and narrower forecast interval.



Model3

Model3 is a forecast model based on ADL, similar to model3. This model utilizes 2 leading indicators, which are high yield spread and building permits. Under AIC, the optimal forecast model includes 7 months lag autoregression of unemployment rate, 6 months lag of high

yield bond spread, and a 1month lag of building permits. This model shows a similar forecast to model 2. However, model3 expects the unemployment rate to return back to 7 percent after a short steep increase in the unemployment rate while model2 expects the unemployment rate to continue increasing.

Conclusion

In conclusion, as each model shows clearly different characteristics, it is not an easy decision to pick one model that forecasts the future unemployment rate in the San Francisco Bay area the best. From model 2,3, it is evident that constructing a forecast model with more variables will have higher variance. Therefore picking the correct variables, after fully understanding the characteristics of the data, that can help forecast the model is crucial when constructing a forecast model. I choose model 2 as my final forecast of the San Francisco MSA unemployment rate.