Predicting US Unemployment Rate Based on Housing Supply Rate and Interest Rate

STATS 170

Robin Lee (7415), Ellen Wei (6652), Xuxin Zhang (7281)

Division of Work:

All members participated in Part I, VI, X

Robin Lee: Part IV, VII

Ellen Wei: Part V, VIII

Xuxin Zhang: Part II, III, IX

I. Introduction

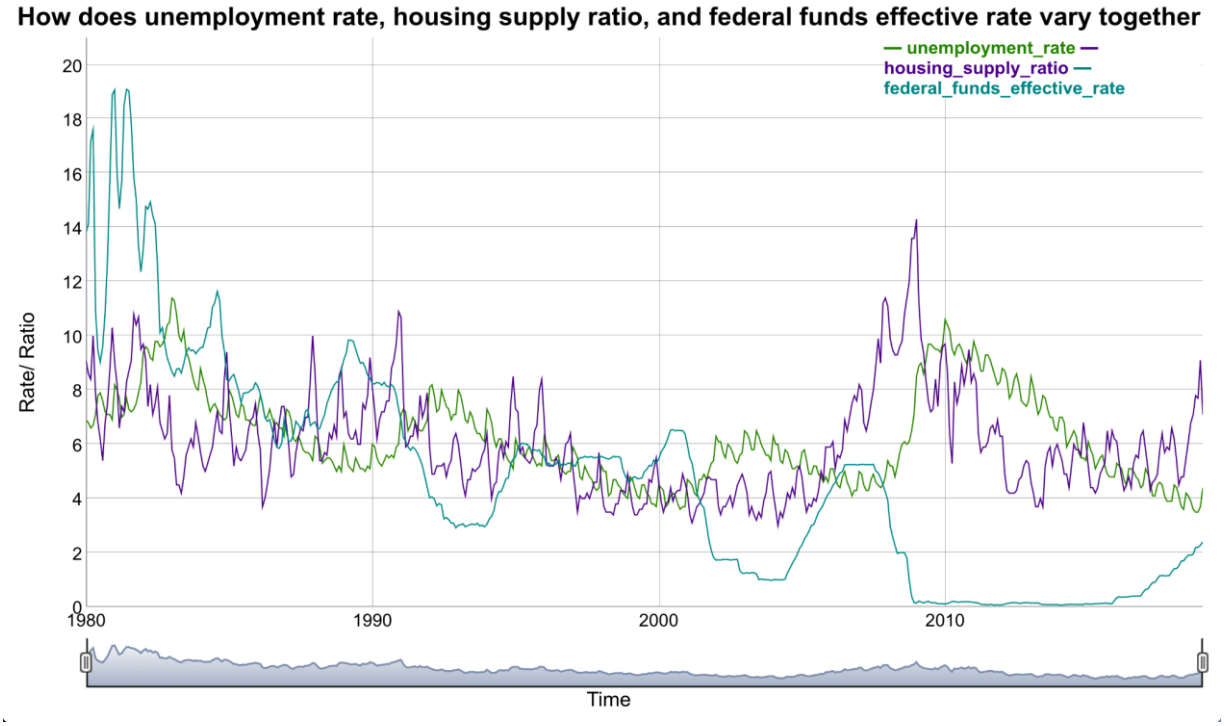
With the number of layoffs in the tech industry escalating throughout the United States following a nation-wide increase in interest rates, we have hypothesized that interest rate is definitely a significant factor that noticeably affects the unemployment rate—raising the question, what other circumstances influence employment? Thus, by applying time series analysis, we attempt to answer the following: what are other potential factors that impact the unemployment rate?

In this research assignment, in addition to interest rate, we have picked the monthly supply of new houses as another independent variable. The intuition behind this choice is that we believe when a massive layoff happens—such as the one recently—people tend to sell their real estate, leading to a significant increase in the housing supply. This is especially true and exemplified during the 2008 Financial Crisis. As such, we are curious as to whether or not this phenomenon is a common pattern. We have chosen three time series datasets: interest rate, unemployment, and housing supply. Below is a table summarizing the time series datasets we have selected from FRED. Notice that all of the data we picked here are **not seasonally adjusted**.

Variable Code Name	Description of what variable measures, frequency, source	Time Period Written	Dependent/ Independent
MSACSRNSA Supply of New Housing in the US	U.S. Census Bureau, Monthly https://fred.stlouisfed.org/series/MSACSRNSA The months' supply is the ratio of new houses for sale to new houses sold.	1963:1-2022:12	Independent
UNRATENSA Unemployment Rate, US	U.S. Bureau of Labor Statistics, Monthly https://fred.stlouisfed.org/series/UNRATENSA 16+ years age, reside in 50 states or District of Columbia, do not reside in institutions or are on active duty in the Armed Forces.	1948:1-2022:12	Dependent
FEDFUNDS Federal Funds Effective Rate	Board of Governors of the Federal Reserve System (US), Monthly https://fred.stlouisfed.org/series/FEDFUNDS Interest rate at which depository institutions trade federal funds (balances held at Federal Reserve Banks) with each other overnight.	1954:7-2022:12	Independent

Table 1: Data Summary

After importing our data into R using the Quandl API, we can display the three time series data on the same plot, as shown in Plot 1. To do this, we first need to make sure that they have the same start time and end time. For simplicity, we picked the time range from 1980-01-01 to 2022-12-01 for all of the three datasets. In this way, we are able to view our full dataset. Later, we will use part of the full dataset as the training data.



Plot 1: How does unemployment rate vary together with housing supply and interest rate

From Plot 1, we can see that the unemployment rate has a similar trend compared to both federal funds effective rate and housing supply ratio. To conduct analysis in later parts of this homework, we have extracted the last 12 data points as the testing data and used the remaining data points as the training data. That is to say, we use the data from 1980-01 to 2018-01 as the training data and that from 2018-02 to 2019-1 as the testing data.

II. Components Features of the Dependent Variable in Training Set

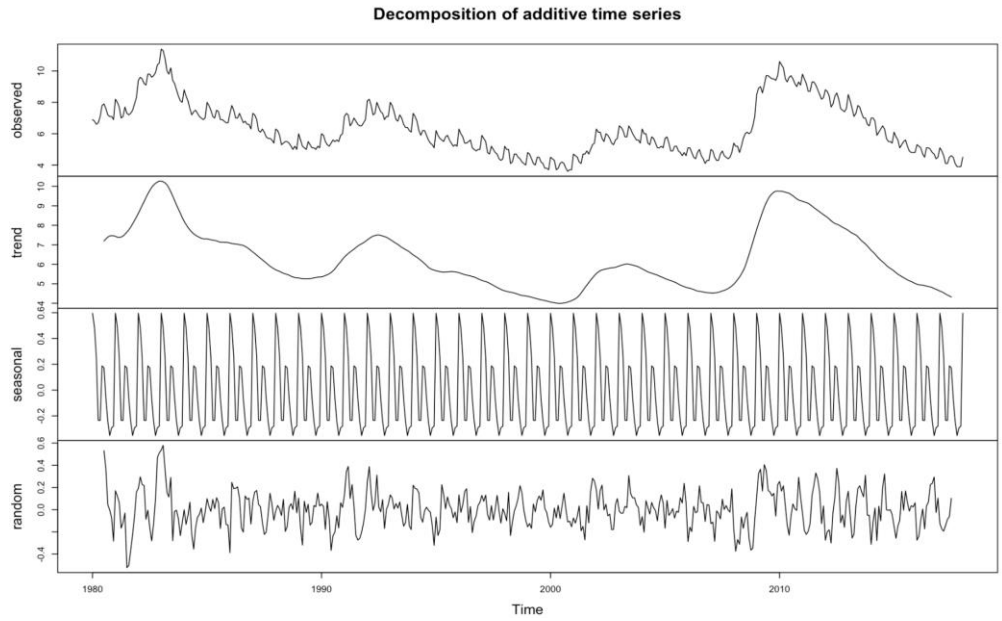
(1) *Decomposition*

Because we want to see how new housing supply and interest rate will affect unemployment, we have chosen **Unemployment Rate** as our dependent variable. Based on the dyplot shown earlier (Plot 1), we can see that there is seasonality and no obvious trend in the unemployment data, as

indicated by the green line. That is to say, we know the seasonality does not increase with the trend. As a result, we can apply additive decomposition, instead of multiplicative decomposition, for our time series.

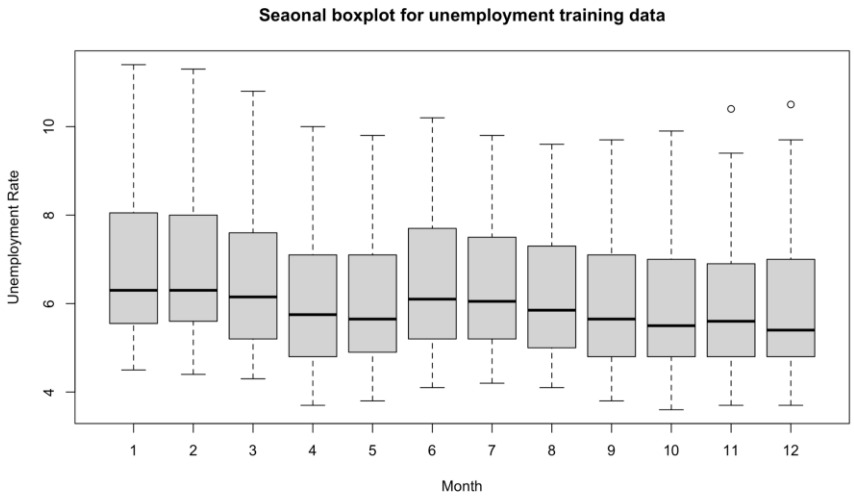
We first plot the additive decomposition of the training data. We see that the range of fluctuation for the random term is relatively small (from -0.4 to 0.4), implying we do not need any transformation on the raw dataset; as such, an additive decomposition is good enough.

.



Plot 2: Additive Decomposition of Unemployment Training Data

(2) *Seasonal Box Plot*

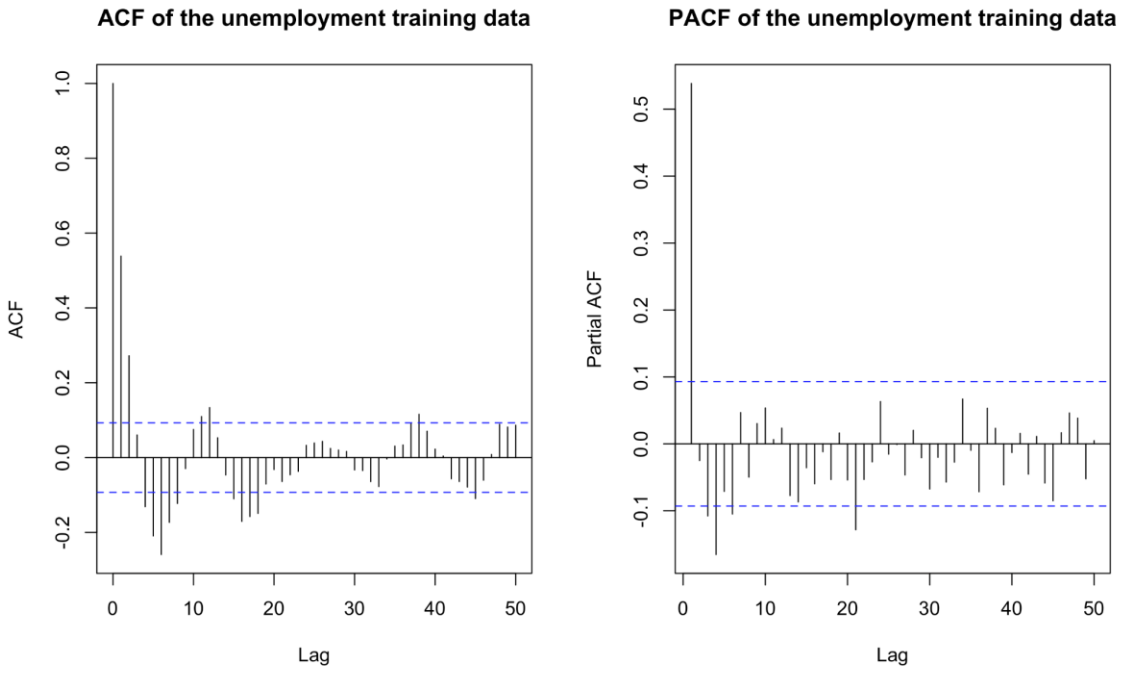


Plot 3: Seasonal boxplot for the unemployment data

From the boxplot (Plot 3), we can see that the unemployment data does show certain seasonality. In particular, we can see that the unemployment rate drops from January to May, before increasing again in June. However, after this growth in June, the unemployment rate begins to drop again from July to December.

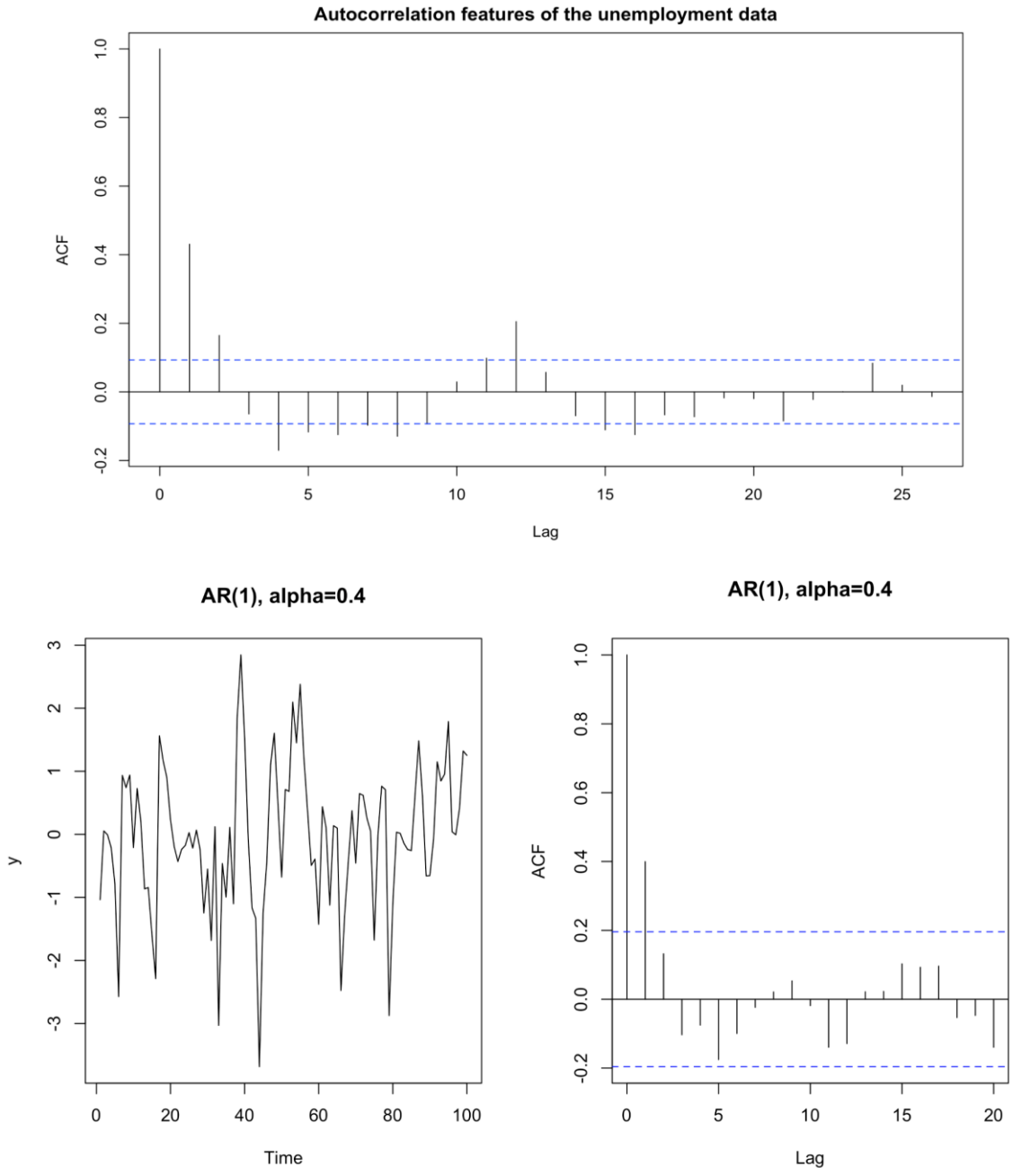
III. Autocorrelation Features of the Dependent Variable Unemployment in the Training Set

Next, we plot the ACF graph for the training data. Looking at the ACF graph (Plot 4), we notice that the autocorrelations for the random term stay significant until a larger lag, but it gradually cuts off. For the regular part, it seems like an ARMA(4, 2) model. The reason is that by looking at the ACF graph, the peak cuts off at lag 2 (MA(2)), and in the PACF, the peak cuts off at lag 4 (AR(4)). For the seasonal part, it looks like the MA(3) model.



Plot 4: ACF and PACF plots for the unemployment data

By comparing the raw time series plots and the ACF plots, we think an autoregressive process with an order of 1 with a parameter of 0.4 might generate our dependent variables; this is because the ACF graphs are similar.



Plot 5: Comparison between ACF of our data and the AR(1) model with parameter of 0.4

IV. Exponential Smoothing Modeling and Forecasting

Using the raw training dependent variable—Unemployment—we fit an appropriate exponential smoothing model and forecast.

As such, we decide to use seasonal Holt-Winters exponential smoothing, which specifically fits a time series with seasonality and trend. Furthermore, we implement additive decomposition, because the model's seasonality does not increase with the trend. The additive Holt-Winters prediction function for time series with period length *p* is defined as:

$$Yhat[t+h] = a[t] + h * b[t] + s[t + 1 + (h - 1) \bmod p]$$

This function tries to find the optimal values of alpha, beta, and/or gamma. We obtain the following results:

Smoothing Parameters:

Alpha:	0.8428519
Beta:	0.02833799
Gamma:	1

Coefficients:

a	4.07894326	**s6**	0.32706991
b	-0.04468121	**s7**	0.01516064
s1	0.40741155	**s8**	-0.40081732
s2	0.25960977	**s9**	-0.52094787
s3	-0.21651062	**s10**	-0.52039932
s4	-0.12115006	**s11**	-0.35669888
s5	0.30746959	**s12**	0.42105674

We forecast as many periods ahead as observations are in the test set, as exemplified below:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2018		4.44 1674	4.24 9191	3.72 8389	3.77 9068	4.16 3007	4.13 7926	3.78 1335	3.32 0676	3.15 5865	3.11 1732	3.23 0751
2019	3.96 3826											

Table 2: Forecast of Raw Training Dependent Variable

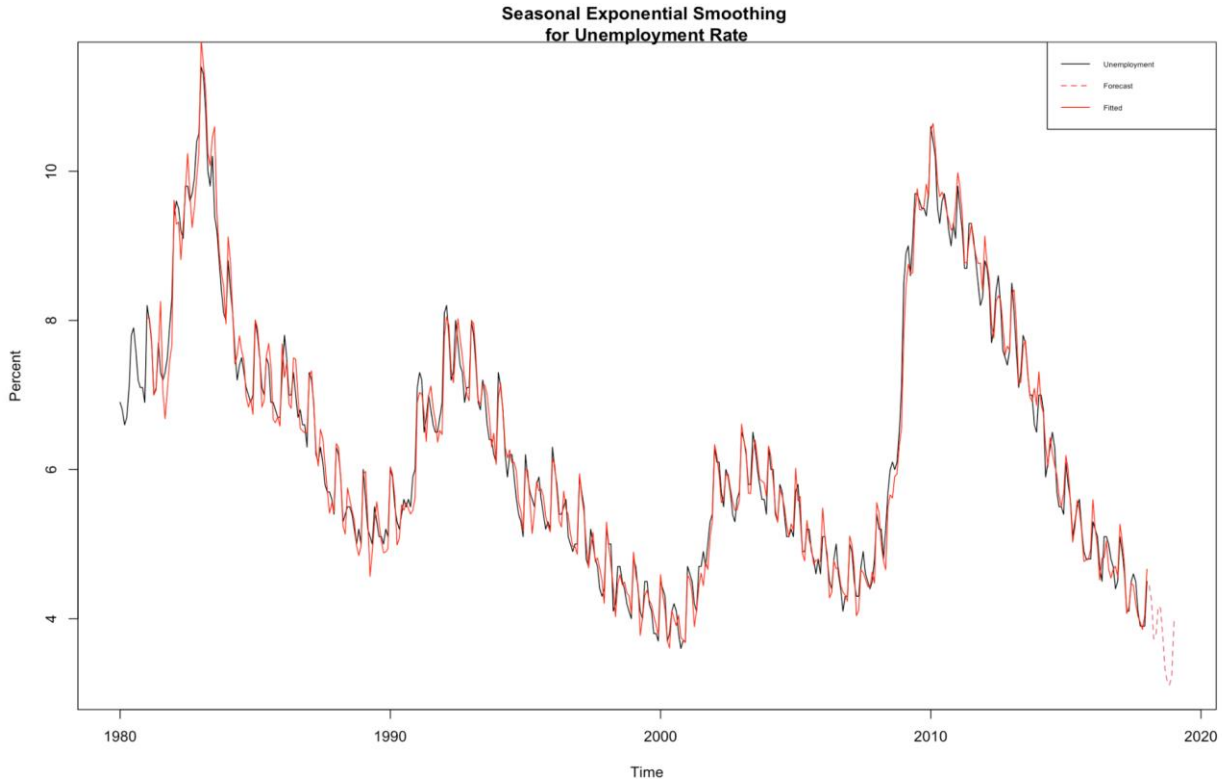
By comparing the test set of the raw data with the forecast we just obtained, we can see that our forecast is fairly accurate at first, but deviates quite significantly after the first six months.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2018		4.4	4.1	3.7	3.6	4.2	4.1	3.9	3.6	3.5	3.5	3.7
2019	4.4											

Table 3: Test Set of Raw Training Dependent Variable

All in all, using additive Holt-Winters on the raw training dependent variable, we forecast the number of unemployed as a percentage of the labor force from February 2018 to January 2019; this is expressed in `forecasts`. We compare these values to `unemployment_rate_test`, which display the true values of the raw data. It appears that the forecast data matches fairly well with the test set of the raw data at first, but grows further apart as the months increase. The forecasted values are about 0.1 from the actual model in the first half of 2018, but the difference increases to more than 0.4 by January 2019.

Finally, we plot the final fitted model and forecast:



Plot 6: Seasonal Exponential Smoothing for Unemployment Rate

The fitted model is defined as:

Yhat[t+h] = (a[t] + h * b[t]) * s[t + 1 + (h - 1) mod p],

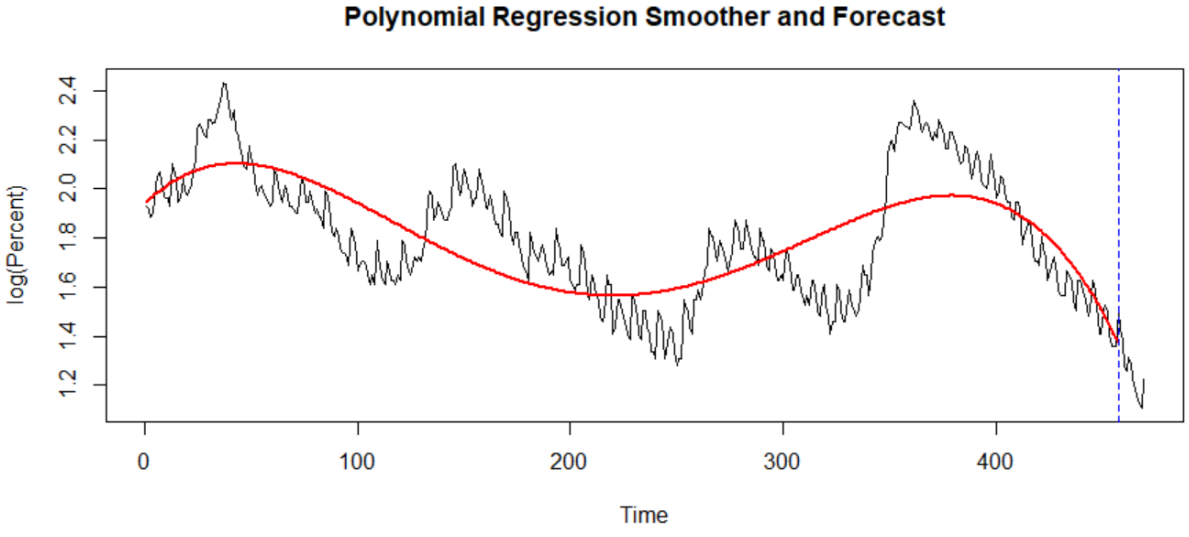
- a[t] = 0.8428519 (Y[t] / s[t-p]) + (1-0.8428519) (a[t-1] + b[t-1])
- b[t] = 0.02833799 (a[t] - a[t-1]) + (1- 0.02833799) b[t-1]
- s[t] = (1) (Y[t] / a[t]) + (1-1) s[t-p]

Using the raw training dependent variable as the primary focus, we can see in the exponential smoothing model (Plot 6) that the forecast predicts that the unemployment rate will continue decreasing throughout 2018 and 2019.

The fitted model (solid red line) appears to fit the actual values (black line) fairly well, and we can see an overall negative trend from 2010 onwards. This signifies that our fitted model is accurate, and that our forecast (dotted line) is most likely accurate as well.

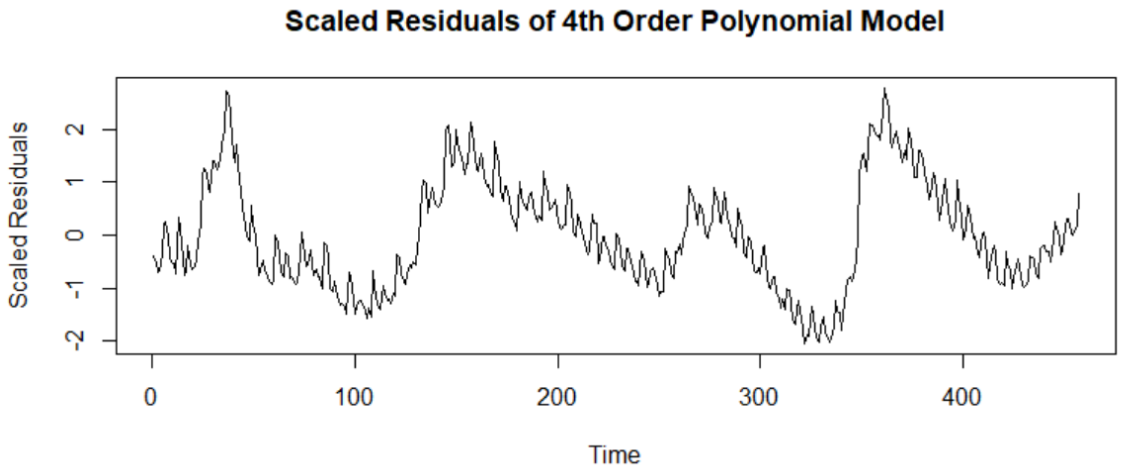
V. Polynomial Regression, Seasonal Effect Modeling, Forecasting

The 4th order polynomial model was selected after fitting 1st, 2nd, 3rd, 4th, and 5th order polynomial models, and assessing their fit to the data. The 1st order (linear) to 3rd order polynomial models were underfitting the data while the 5th order polynomial appeared to be overfitting the data.



Plot 7: Polynomial Regression Smoother and Forecast

Plot 7 shows the plot of a 4th order polynomial model fitted to the training data. The red line represents the trend as modeled by polynomial smoother. The blue dotted line indicates where the forecast begins, with seasonality added to the trend.



Plot 8: Scaled Residuals of 4th Order Polynomial Model

We determine the goodness of fit of a regression model with a standardized time plot of the residuals (Plot 8). Unfortunately, all of the polynomial models are not great models for the unemployment rate data as we can observe cyclical patterns in Plot 8.

Table 4 displays the additive decomposition with 4th order polynomial regression- $\hat{T}$ are the predicted trend values, $\hat{S}$ are the predicted seasonal values, and $\hat{x}$ is the predicted value.

Date	t	$\hat{T}$	$\hat{S}$	$\hat{x}$
2018, 2	458	1.357600	0.07638777	1.433987
2018, 3	459	1.338896	0.04099895	1.379895
2018, 4	460	1.319833	0.03898064	1.280852
2018, 5	461	1.300407	0.03921716	1.261189
2018, 6	462	1.280614	0.03157355	1.312187
2018, 7	463	1.260450	0.03124312	1.291693
2018, 8	464	1.239913	0.00749424	1.232419
2018, 9	465	1.218999	0.03701035	1.181989
2018, 10	466	1.197705	0.05767968	1.140025

2018, 11	467	1.176025	0.04700606	1.129019
2018, 12	468	1.153958	0.04676550	1.107193
2019, 1	469	1.131499	0.09395024	1.225450

Table 4: Additive decomposition with polynomial regression

Equation of fitted polynomial regression model:

$$y_t = 1.938 + 8.454(10^{-3})t - 1.287(10^{-4})t^2 + 5.054(10^{-7})t^3 - 5.910(10^{-10})t^4$$

VI. Conclusion

In conclusion, we found in the time series that the dependent variable—unemployment rate—contains the seasonality that does not increase with the trend. These two features indicate that the seasonality of unemployment does not increase with the trend, allowing us to utilize additive decomposition rather than multiplicative decomposition. Hence, after plotting the additive decomposition outcomes, it is clear that the random term fluctuates around the value of 0—signifying that our data has constant variance.

Using seasonal Holt-Winters exponential smoothing, we were able to relatively accurately forecast data in 2019 by constructing a training dataset containing data before 2019, and comparing it with a test dataset containing only 2019 data. As such, the comparison appears to show that the forecast data matches fairly well with the test set of the raw data.

The RMSE for the exponential smoothing model is **0.2632173**, the RMSE for the polynomial regression model is **1.222444**, and the average forecast RMSE is **0.717327**(Table 3). We can say that the exponential smoothing model performed the best so far in fitting the unemployment rate data and the polynomial regression model the worst so far.

Date	**Raw Data Values**	**Exponential Smoothing**	**Polynomial Regression**	**Average Forecast**
2018, 2	4.4	4.441674	3.939979	4.190826
2018, 3	4.1	4.249191	3.589400	3.919295
2018, 4	3.7	3.728389	2.977189	3.352789

2018, 5	3.6	3.779068	2.851970	3.315519
2018, 6	4.2	4.163007	3.148521	3.655764
2018, 7	4.1	4.137926	3.003566	3.570746
2018, 8	3.9	3.781335	2.635782	3.208559
2018, 9	3.6	3.320676	2.325321	2.822999
2018, 10	3.5	3.155865	2.078094	2.616979
2018, 11	3.5	3.111732	1.999822	2.555777
2018, 12	3.7	3.230751	1.871225	2.550988
2019, 1	4.4	3.963826	2.602641	3.283233
	RMSE:	**0.2632173**	**1.222444**	**0.717327**

Table 5: Raw data, forecast from models, and RMSE

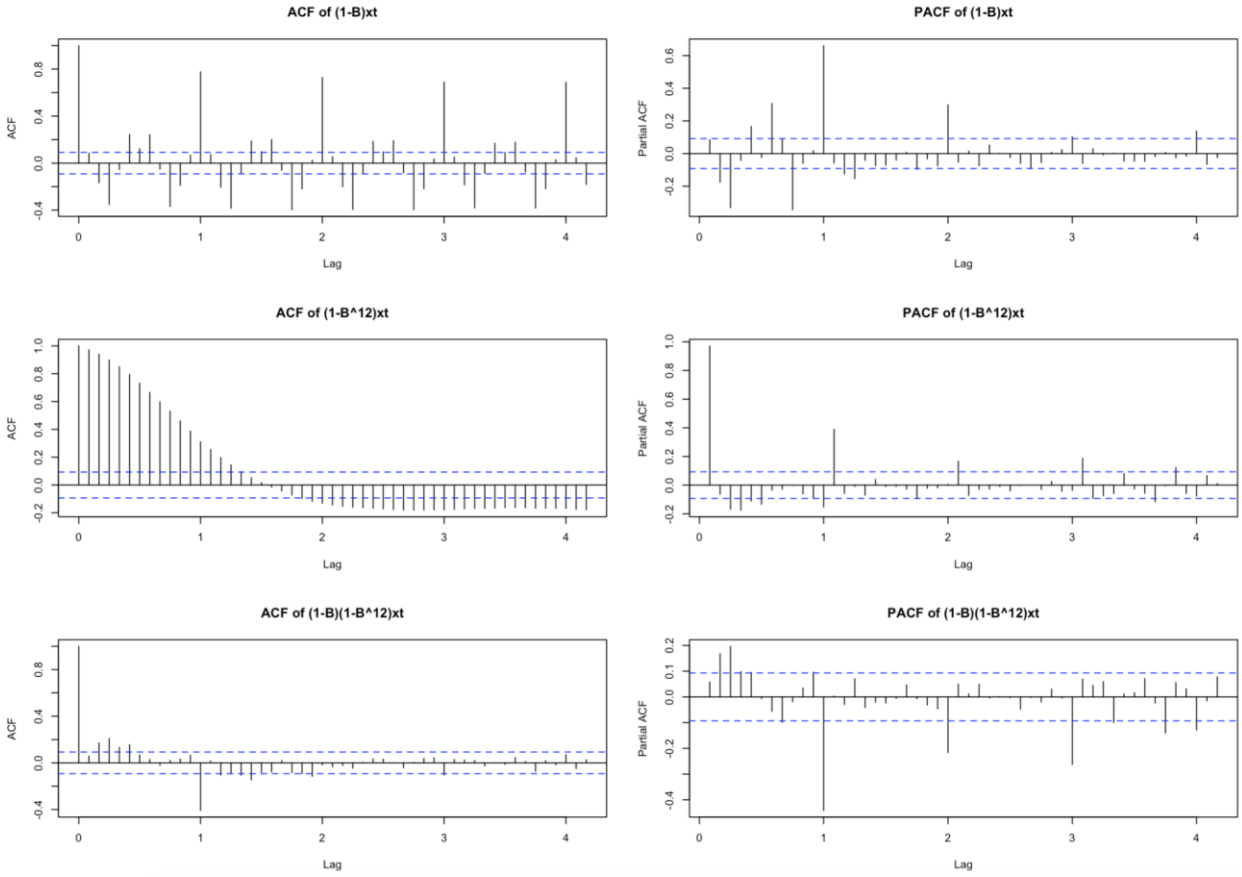
VII. ARIMA Modeling and Forecasting

VII.1 Addressing Pre-Transformations

We have ultimately decided that we do not need to apply any pre-transformations, since we found that additive decomposition works—as indicated in Section II. Because the **variance** in Unemployment Rate is stationary, no transformation is required to ensure the constant variance. Hence, we keep our dependent training data as "y".

VII.2 Assessment of Mean Stationarity

As exemplified in Plot 4: ACF and PACF plots for the unemployment data, there is seasonality and trend in the mean when we implement a lag of 50 in our ACF and PACF models. In the ACF and PACF of Regular and Seasonal Differencing (Plot 9), we analyze the ACF and PACF of the model alongside regular and/or seasonal differencing in order to identify what method minimizes trend in the mean and make the series stationary.



Plot 9: ACF and PACF of Regular Difference, Seasonal Difference, and Seasonal Difference of the Regular Difference

By seasonally differencing the regular differencing, we notice that the ACF has few significant spikes at the smaller lags before cutting off. Even though we identify a significant spike at lag = 1, we see that the lag quickly dies off. Therefore, we conclude that with the seasonal differencing of the regular differencing of the training data, we achieve mean stationarity.

VII.3. Model Identification

For model selection, we break the ACF and PACF of seasonal difference of regular difference into regular and seasonal parts. From the ACF and PACF of regular and seasonal differencing (Plot 9), we see that the ACF cuts off at approximately lag 1, and the PACF cuts off at approximately lag 3. Because there is seasonality, we need to identify a model for both the stationarity at low lags and the stationarity at seasonal lags.

As we mentioned, the significant peak cuts off after the third spike in the PACF plot, which tells us that it corresponds to an AR(3). For the seasonal part, we see that the significant spike dies off in both ACF and PACF plots, so we conclude this is an ARMA(1, 1).

Therefore, we have decided that the regular differencing can be expressed as an AR(3) and MA(1), while the seasonal differencing can be expressed as AR(1) and MA(1). The model—in ARIMA notation ARIMA(p,d,q)(P,D,Q)_F—is:

ARIMA(3,1,1)(1,1,1)_12

VII.4 Fit and Diagnose

In order to fit and diagnose our identified model, we apply the Ljung-Box test using 50 lags—as seen in the ACF (Plot 9). Given the following hypotheses,

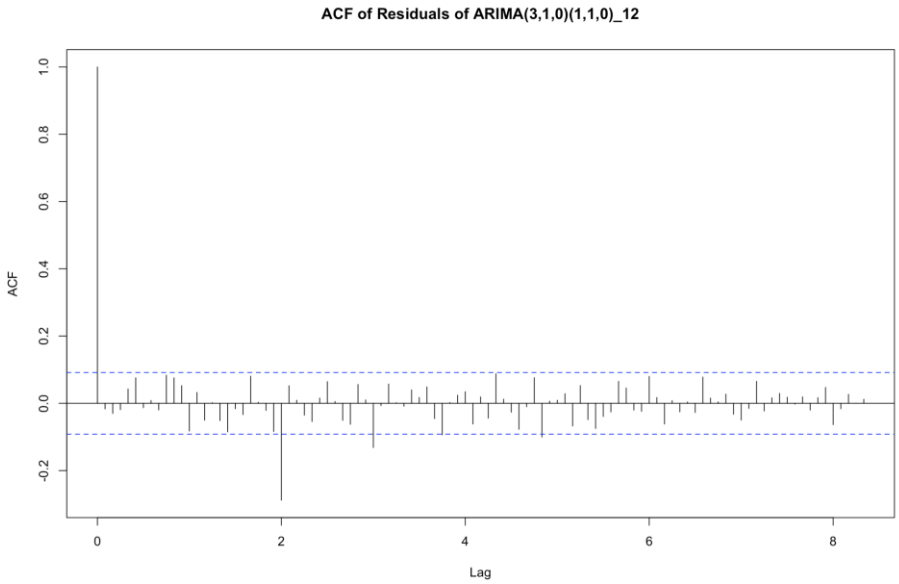
H_0: rho_1=rho_2=...=rho_50=0
H_a: not all rho_k up to lag k are 0

We want to test whether or not our detrended and seasonally adjusted Unemployment data are white noise. By conducting this test given alpha=0.05, we obtain: **P-value: 0.82536**

As such, because of the P-value > alpha, we do not reject the null hypothesis that the residuals are white noise. This is good, because the residuals of a well-fitting model should be white noise. Therefore, by the Ljung-Box test, **we do not need to restart the model fitting process.**

Furthermore, we fit two more models and compare them with our model to see which one is better by analyzing their respective AIC values. The best model is the one with the smallest AIC. The other two models we compare are ARIMA(3,1,0)(1,1,0)_12 and ARIMA(5,1,0)(1,1,0)_12.

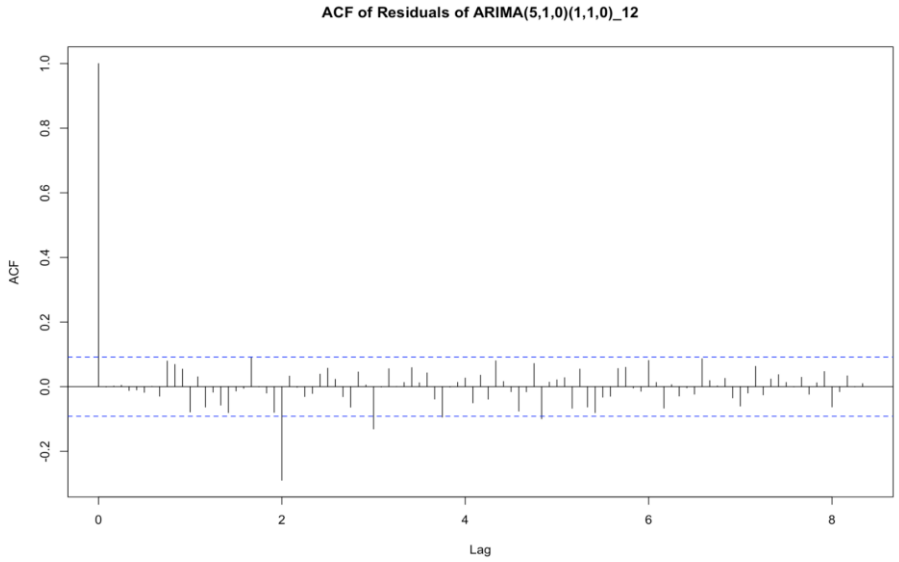
- **Model 1:** ARIMA(3,1,0)(1,1,0)_12



Plot 10: ACF of Model 1: ARIMA(3,1,0)(1,1,0)_12

ARIMA(3,1,0)(1,1,0)_12 AIC = **-130.2859**

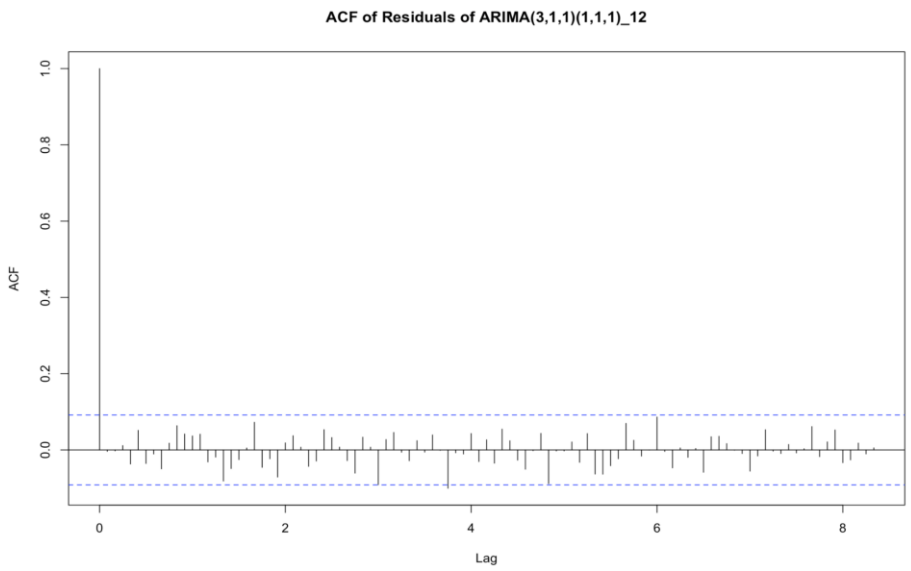
- **Model 2:** ARIMA(5,1,0)(1,1,0)_12



Plot 11: ACF of Model 1: ARIMA(5,1,0)(1,1,0)_12

ARIMA(5,1,0)(1,1,0)_12 AIC = **-131.879**

- **Model 3:** ARIMA(3,1,1)(1,1,1)_12



Plot 12: ACF of Model 1: ARIMA(3,1,1)(1,1,1)_12

ARIMA(3,1,1)(1,1,1)_12 AIC = **-237.1305**

Therefore, the best model is ARIMA(3,1,1)(1,1,1)_12, because it has the lowest AIC.

The model coefficients are:

Coefficients of ARIMA(3,1,1)(1,1,1)_12:

	ar1	**ar2**	**ar3**	**ma1**	**ma2**	**sma1**
	0.6052	0.1334	0.0921	-0.5824	0.0495	-0.8786
se	0.1353	0.0558	0.0669	0.1293	0.0570	0.0285

Hence, our model in the polynomial form—with all AR terms and differencing on the left hand side of the = sign and all the MA terms on the right hand side—is:

(1-0.6052B-0.1334B^2-0.0921B^3)(1-B^12)(1-B)y_t = (1-0.5824B+0.0495B^2)(1-0.8786B^12)w_t

Expanding the polynomial version of our mode, the final forecasting equation, with all the independent variables on the right hand side and only the dependent variable at time t on the left-hand-side, is:

(1-0.6052B-0.1334B^2-0.0921B^3)(1-B^12)(1-B)y_t = (1-0.5824B+0.0495B^2)(1-0.8786B^12)w_t

⇒ y_t (-0.0921B^16 - 0.0413B^15 - 0.4718B^14 + 1.6052B^13 - B^12 + 0.0921B^4 + 0.0413B^3 + 0.4718B^2 – 1.6052B+1) = w_t (1 – 0.8786B^12 – 0.5824B + 0.51169664B^13 + 0.0495B^2 – 0.0434907B^14)

⇒ -0.0921y_(t-16) - 0.0413y_(t-15) - 0.4718y_(t-14) + 1.6052y_(t-13) - y_(t-12) + 0.0921y_(t–4) + 0.0413y_(t–3) + 0.4718y_(t–2) – 1.6052y_(t–1)+y_t = w_t – 0.8786w_(t–12) – 0.5824w_(t–1) + 0.51169664w_(t–13) + 0.0495w_(t–2) – 0.0434907w_(t–14)

⇒ **y_t = 0.0921y_(t-16) + 0.0413y_(t-15) + 0.4718y_(t-14) - 1.6052y_(t-13) + y_(t-12) - 0.0921y_(t-4) - 0.0413y_(t-3) - 0.4718y_(t-2) + 1.6052y_(t-1) - 0.8786w_(t-12) - 0.5824w_(t-1) + 0.51169664w_(t-13) + 0.0495w_(t-2) - 0.0434907w_(t-14)**

We want to check whether the model is stationary and invertible, so we find the roots of the MA and AR parts:

- *Stationarity (AR roots):*
  Finding the modulus of the roots of the polynomial in B: (1,-0.6052,-0.1334,-0.0921)
  ⇒ 1.140201 3.085883 3.085883
  Because the modulus of the roots are greater than 1, the process is stationary.
- *Invertibility (MA roots):*
  Finding the modulus of the roots of the polynomial in B: (1, -0.5824, 0.0495)
  ⇒ 2.0873509 9.6783056
  Because the modulus of the roots are greater than 1, the process is invertible.

In order to confirm that the coefficients generating the data are not 0, we utilize t-tests. Our hypothesis are:

H0: alpha_1, alpha_2, alpha_3, alpha_4, alpha_5, alpha_6 = 0
Ha: alpha_1, alpha_2, alpha_3, alpha_4, alpha_5, and/or alpha_6 not equal to 0

	alpha_1	alpha_2	alpha_3	alpha_4	alpha_5	alpha_6
value	0.6052	0.1334	0.0921	-0.5824	0.0495	-0.8786
se	0.1353	0.0558	0.0669	0.1293	0.0570	0.0285
t_(n-k)	4.47302	2.39068	1.37668	4.50425	0.86842	30.82807

We reject the null hypotheses in all cases, because the t-statistics are more than 2 standard errors away from the center. Thus, there is statistically evidence suggesting that the coefficients of the AR and MA models generating the data—labeled alpha—are not 0.

VII.5 Forecasting

After confirming that the residuals are white noise, the model is stationary and invertible, and the model coefficients are significantly different from 0, we proceed to use the model we have selected in VII.3 to forecast the future values of the series.

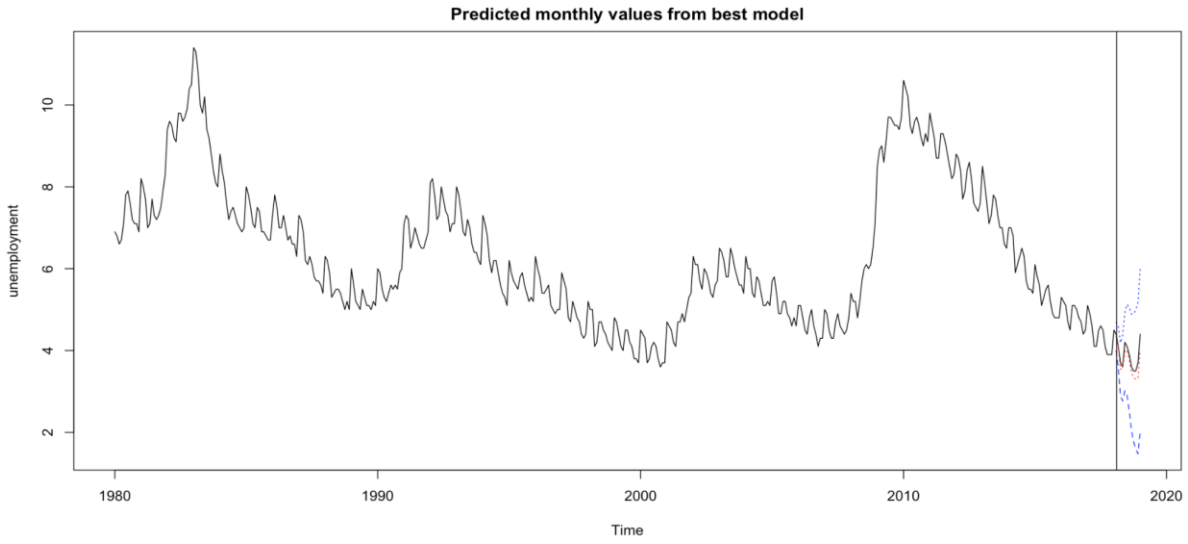
First and foremost, we construct confidence intervals that inform us as to how confident we are that the actual value of the series in the future lies in the particular interval. Thus, with our final model—ARIMA(3,1,1)(1,1,1)_12—containing normally-distributed residuals, we forecast the test period of our original raw y variable (Unemployment) to obtain point and prediction intervals. By doing so, we print the data frame with raw test values from Unemployment, the forecast, the forecast interval (CI Low, CI High) and standard error of our forecast.

Test Values	CI Low	Forecast Value	CI High	Forecast SE
4.4	3.9885357	4.3393331	4.6901305	0.17897825
4.1	3.5827438	4.0845305	4.5863172	0.25601362
3.7	2.8913789	3.5396347	4.1878905	0.33074274
3.6	2.7654800	3.5691688	4.3728577	0.41004533
4.2	3.0167901	3.9750120	4.9332340	0.48888873
4.1	2.9065643	4.0181760	5.1297878	0.56714886
3.9	2.5231823	3.7867982	5.0504140	0.64470195
3.6	2.0609218	3.4743871	4.8878525	0.72115578
3.5	1.7703723	3.3311271	4.8918819	0.79630347
3.5	1.5895574	3.2947801	5.0000029	0.87001161
3.7	1.4740907	3.3207743	5.1674579	0.94218553
4.4	2.0286240	4.0136551	5.9986862	1.01277097

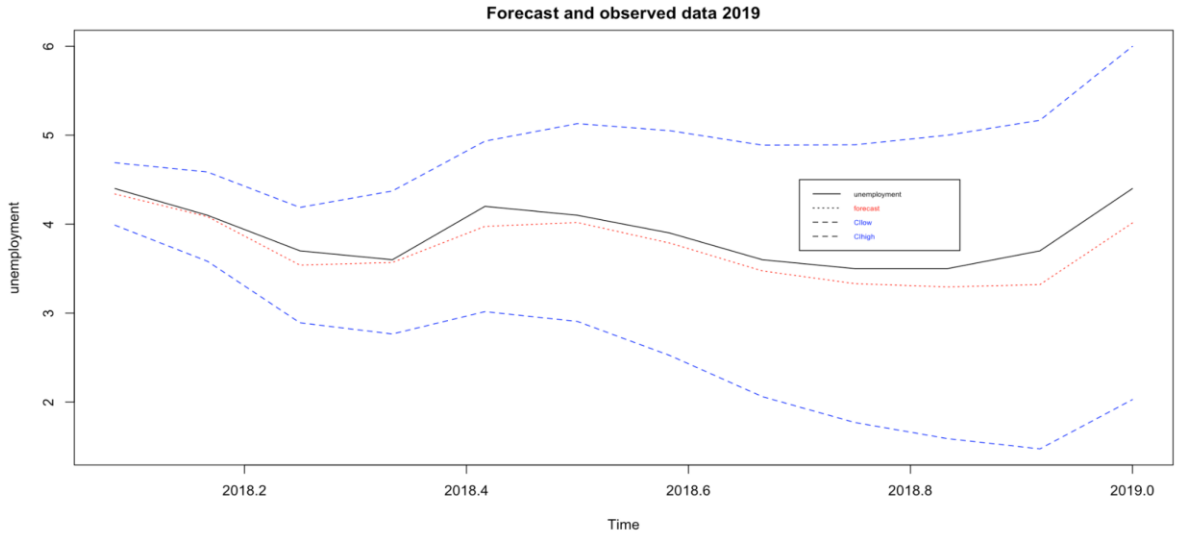
Table 6: Forecast of ARIMA(3,1,1)(1,1,1)_12

We can see that the forecasted values tend to be rather close to the raw test values, and are all within 1 SE of the actual value. Furthermore, the test values are within the prediction interval. As with the exponential smoothing model, as the months increase, the forecast grows farther from

the test value; however, the actual value of the series in the future still lies in the particular interval. In conclusion, ARIMA(3,1,1)(1,1,1)_12 provides an accurate forecast of the test data.



Plot 13: Entire Forecast using ARIMA(3,1,1)(1,1,1)_12

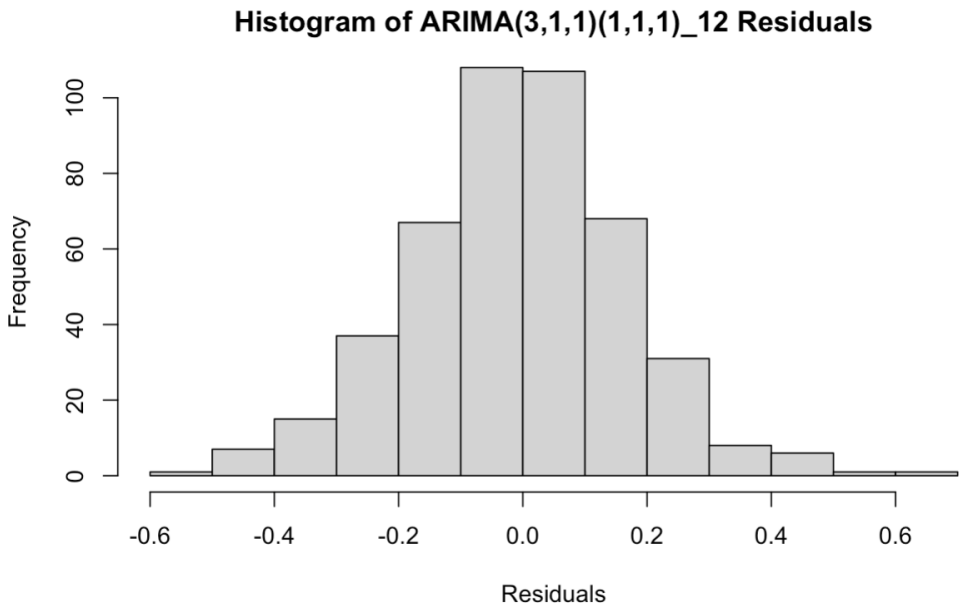


Plot 14: Forecast Period using ARIMA(3,1,1)(1,1,1)_12

Next, as demonstrated in class, we plot the forecast—first with the whole time series, before plotting just the forecast period. From Plot 13 and Plot 14, we see that the forecast (red dotted line) appears to follow the raw unemployment data (black line) fairly closely, with the upper and lower bands (blue dotted lines) providing a sufficient buffer for the prediction.

We measure the accuracy of our forecast using the root mean square error (RMSE) statistic. Our final calculated MSE is: **0.56367369**
Because RMSE values between 0.2 and 0.5 are able to relatively predict the data accurately, we are satisfied with the capabilities of our model—ARIMA(3,1,1)(1,1,1)_12.

Finally, besides the other assumptions we have regarding the residuals w_t having mean a mean of 0, a constant variance at all t, and 0 correlation, we need to check that it is normally distributed in order to conduct accurate statistical inference. Hence, we want to check that the residuals for our final model, ARIMA(3,1,1)(1,1,1)_12, are normally distributed; as such, we plot a histogram (Plot 13) to confirm that this is the case:



Plot 15: Histogram of ARIMA(3,1,1)(1,1,1)_12 Residuals

As we can see in Plot 15, the histogram appears to be normal; we are able to proceed with the forecast we have created.

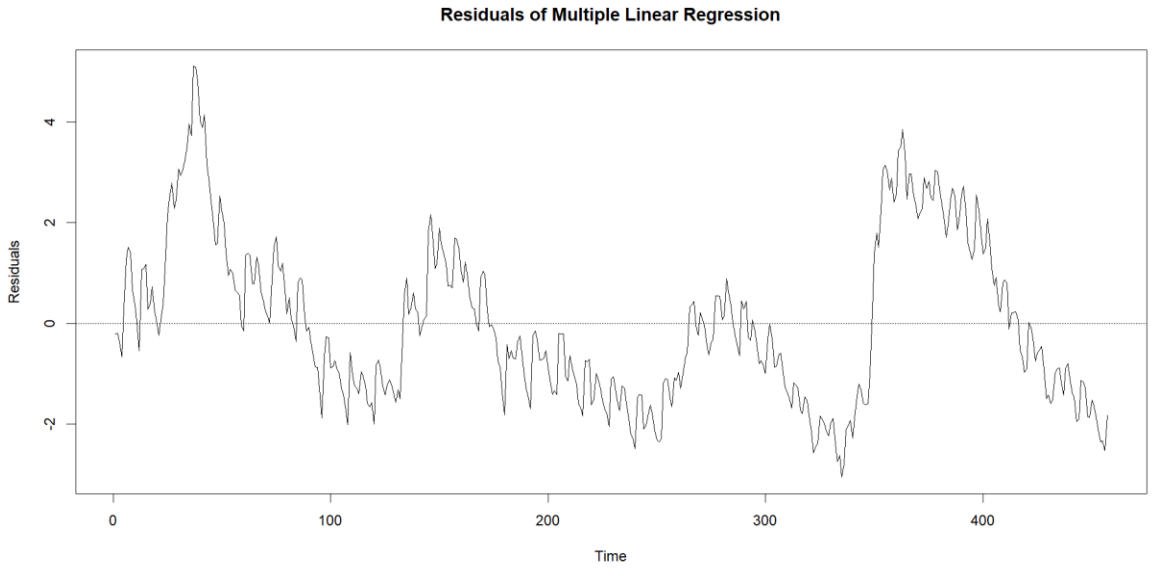
VIII. Multiple Regression with ARMA Residuals

VIII.1 Causal Model Fit

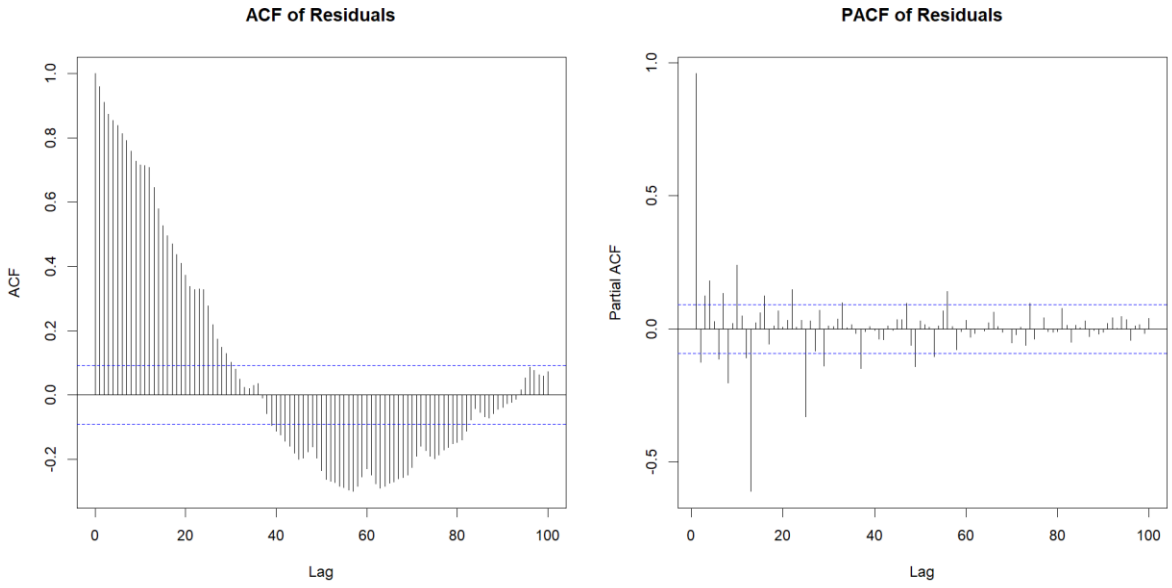
To fit a multiple regression model with our variables, we first split the three time series into training and testing data. After that, we fit a multiple regression model with unemployment rate

as the dependent variable, and housing supply ratio and federal funds effective rate as the two independent variables.

To determine whether the model was a good fit to the data, we plot the residuals (Plot 16).

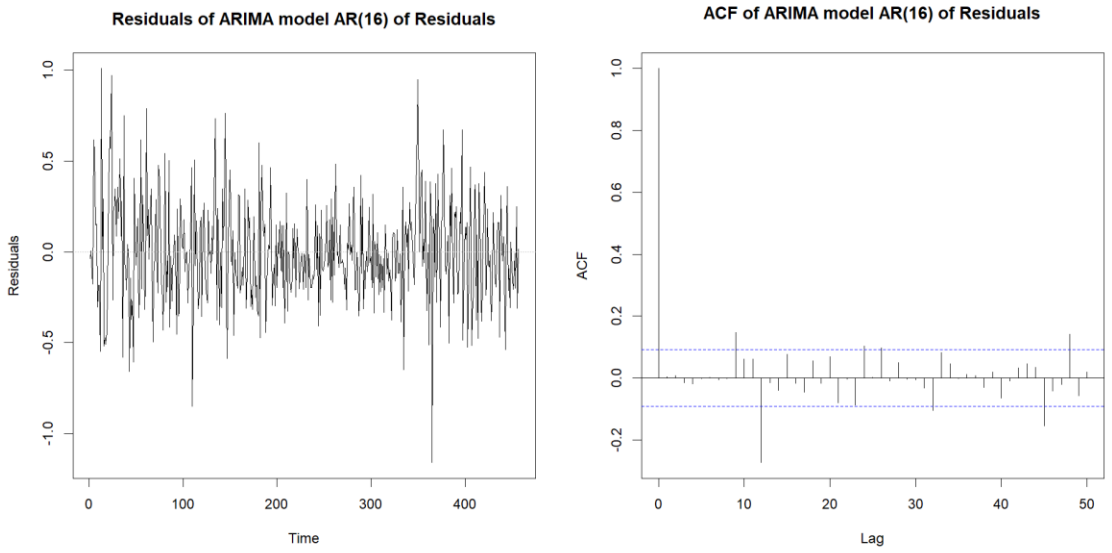


Plot 16: Residuals of multiple linear regression (Unemployment Rate on Housing Supply Ratio and Federal Funds Effective Rate)



Plot 17: ACF and PACF of residuals of multiple linear regression model

We observe a cyclical pattern from the ACF and PACF plots (Plot 17), so we further determine that the residuals are not white noise. Hence, we attempt to fit an ARMA model which we can get the coefficients of to feed into a GLS model.



Plot 18: Residuals plot and ACF plot of residuals of AR(1) model of residuals of multiple linear regression model

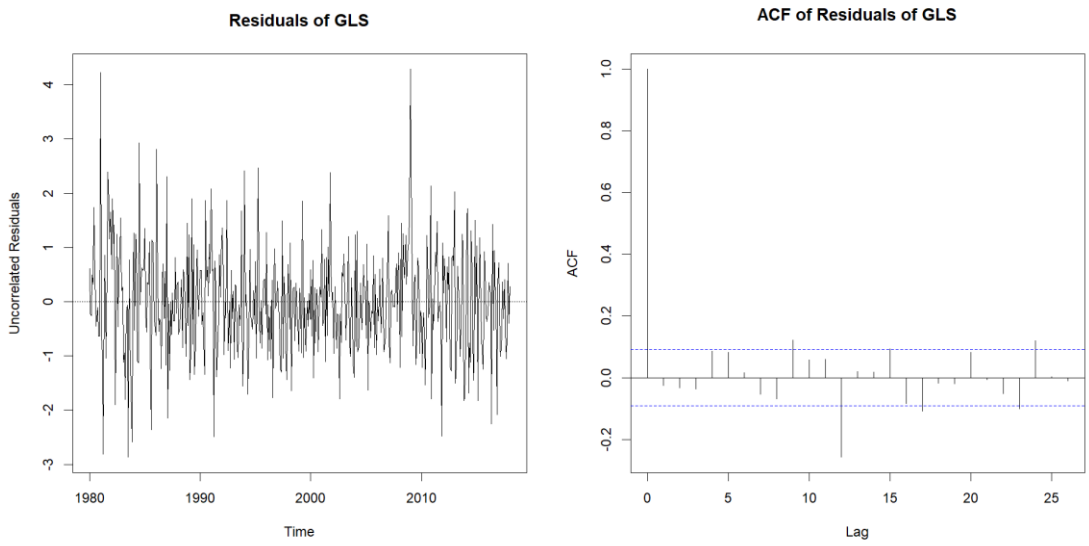
After much trial and error, we selected the ARMA model with the best AIC value, AR(16). We find that this model is stationary when we observe the residuals to be white noise and as shown in the ACF plot (Plot 18).

We then attempted to feed the coefficients we found into a GLS model, but were initially unable as not all the coefficients were less than 1. This meant that our model was not invertible, and hence could not be used for calculating a GLS model.

We repeated this process multiple times with different ARMA models that appeared to have white noise residuals, and could not determine a set of coefficients that were all less than 1.

Ultimately, we only used the fact that the AR model was of order 16 as a parameter in the GLS function in R without inputting our own coefficients.

The GLS function was able to calculate its own coefficients. Plot 19 displays the residuals of the GLS model and the ACF of the GLS residuals.



Plot 19: Residuals and ACF plot of GLS model

The formula that we found is the following where x_{1t} is housing supply ratio at time t and x_{2t} is federal funds effective rate at time t:

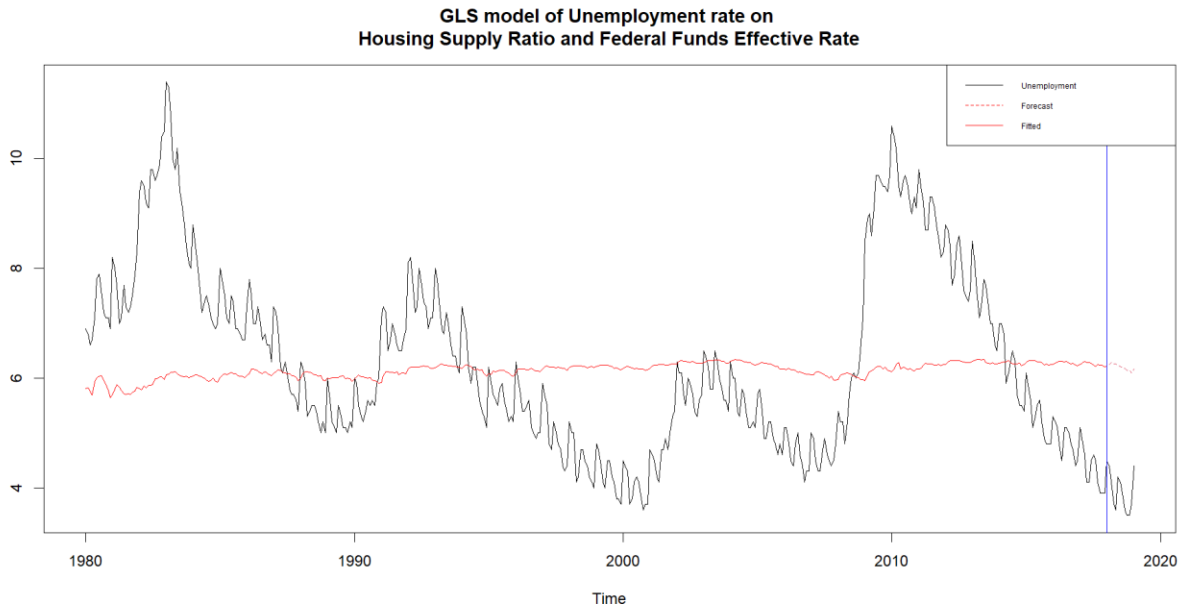
$$y_t = 6.480278 - 0.036687x_{1t} - 0.024086x_{2t} + e_t$$

$$e_t = 0.947691454e_{t-1} + 0.136519195e_{t-2} + 0.005222809e_{t-3} - 0.064356943e_{t-4} + 0.048716685e_{t-5} - 0.133601897e_{t-6} + 0.101303293e_{t-7} - 0.027352147e_{t-8} - 0.149792849e_{t-9} + 0.089271534e_{t-10} + 0.035347094e_{t-11} + 0.687864584e_{t-12} - 0.643179964e_{t-13} - 0.142164182e_{t-14} - 0.127062788e_{t-15} + 0.217739528e_{t-16}$$

Intercept standard error: 0.5313121

x_{1t} standard error: 0.0142313
x_{2t} standard error: 0.0125550

Plot 20 shows the plot of the training data and the forecasted values. The RMSE of the GLS model on the test data is **2.328397**.



Plot 20: Plot of Unemployment Rate, GLS fitted values, and forecast

IX. Vector Autoregression

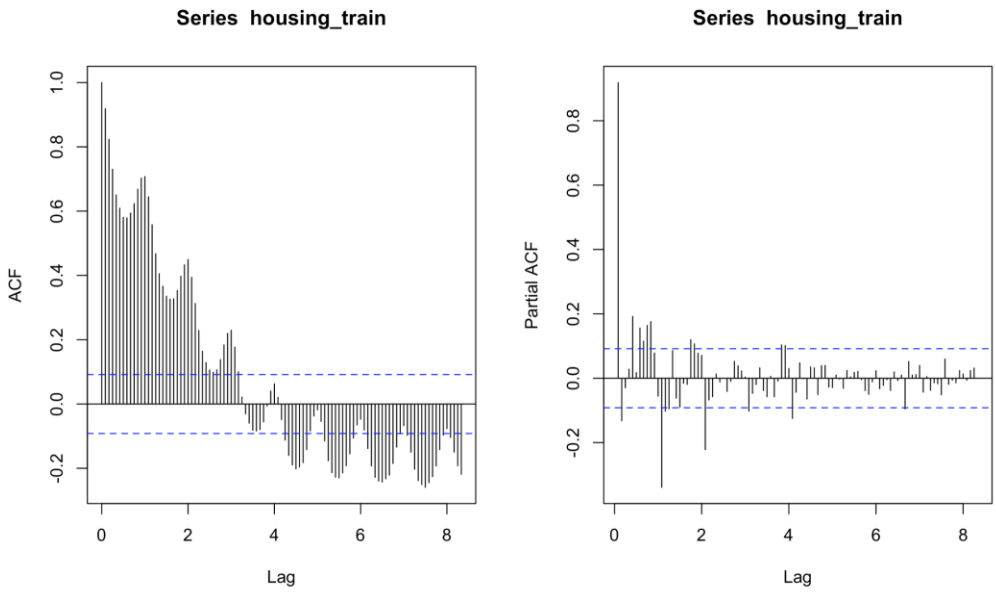
IX.1 CCF and degree for VAR

In order to conduct CCF (cross-correlation function), we need to make sure the time series data we input are stationary. We first draw the ACF and PACF for both of our independent variables: Federal Reserve Effective Rate and Housing Supply Ratio.

However, before that, we need to split the data for independent variables into training and testing sets in order to measure the accuracy of our model. Just as what we did to the dependent variable, for each of the two independent variables, we use the data from 1980-01 to 2018-01 as the training data and that from 2018-02 to 2019-1 as the testing data.

For the unemployment training data, we have already concluded in VII.2 that we will apply seasonal differences and regular differences to it. Thus, $unemployment_t = (1 - B^{12})(1 - B)unemployment_t$.

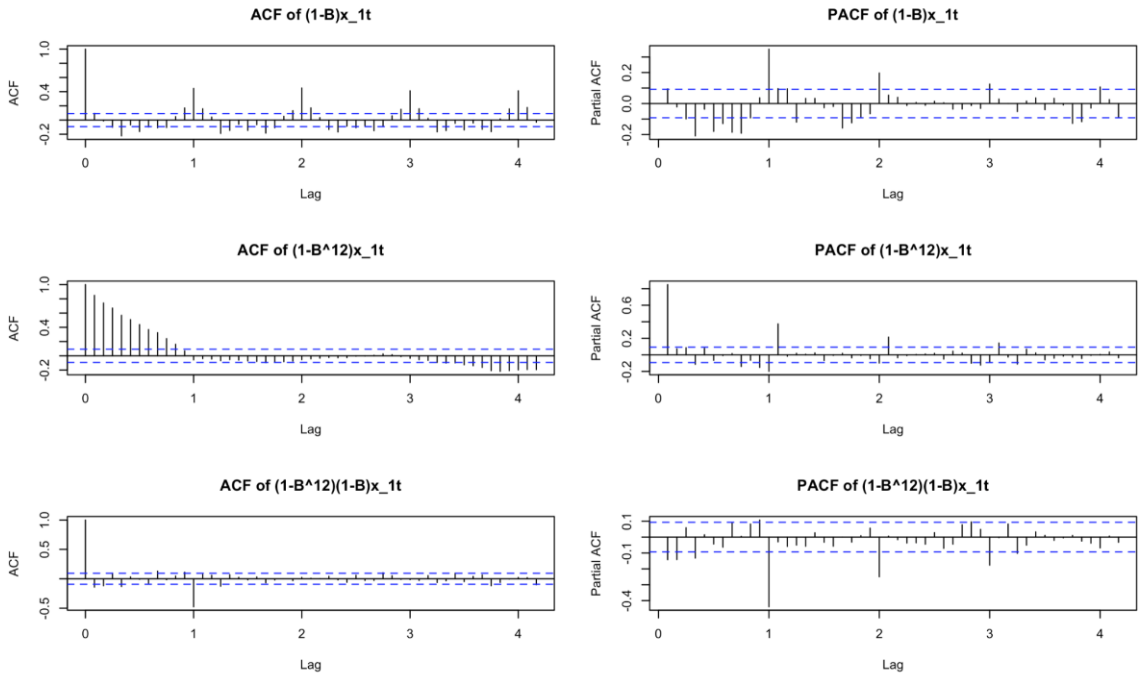
Next, we take a look at the ACF and PACF for the housing supply ratio training data (Plot 21). The data is obviously not mean-stationary.



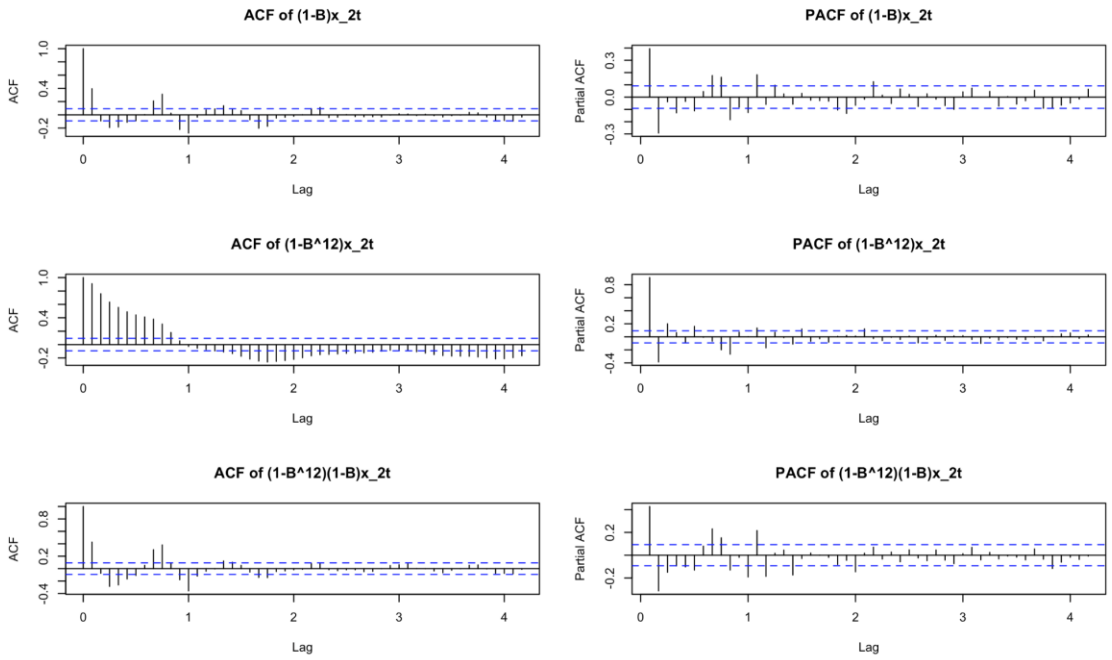
Plot 21: ACF and PACF of Housing Supply Ratio training data

We then try the regular difference, seasonal difference, and the seasonal difference of the regular difference of the training data. We can see from Plot 22 that when we apply seasonal difference to the regular difference to the training data, we achieve stationarity. Thus, for housing supply ratio, we go with:

$housing.supply.ratio_t = (1 - B^{12})(1 - B)housing.supply.ratio_t.$



Plot 22: ACF and PACF of all kinds of differencing applied to Housing Supply Ratio training data



Plot 23: ACF and PACF of Differencing applied to Federal Reserve Effective Rate Training Data

Finally, for the federal reserve effective rate, from Plot 23, we believe it is sufficiently enough just to apply regular differences to achieve stationarity.

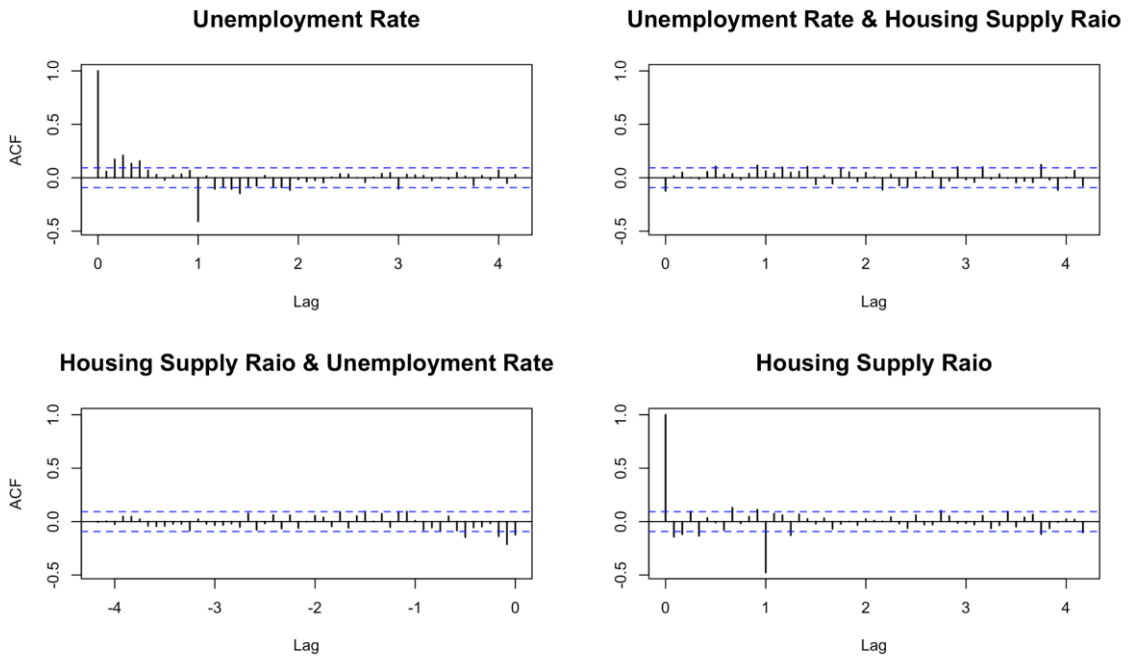
Thus, we obtain: $fed.reserve.rate_t = (1 - B) fed.reserve.rate_t$.

After figuring out the differencing we should apply to the training data of both our dependent and independent variables to make them stationary, next we will study their cross-correlation. Before that, we need to combine all of our differenced data into a single object.

From Plot 24, we see the cross-correlation between unemployment and housing supply ratio. Based on this, we can find the VAR model between them.

Since we care about how housing supply ratio affects unemployment, we look at the top right graph in Plot 24. We see that the first significant spike occurs at lag 6 (excluding lag 0).

In addition, that significant spike decays right away, so we know unemployment at time t depends on housing supply ratio at time (t-6). Next, we see how unemployment is related to itself. We see that CCFs die away in a damped sine-wave fashion, so we know unemployment at time t depends on unemployment at time t-1 and unemployment at time t-2.



Plot 24: CCF between Unemployment and Housing Supply Ratio training data

Writing this in formula form, we have:

$y^*_t = a_1 x^*_{1,t-6} + a_2 y^*_{t-1} + a_3 y^*_{t-2}$.

[*Here, we use* y^*_t *to denote the differenced unemployment training data and use* $x^*_{1,t}$ *to denote differenced housing supply ratio training data*]

When we look at how the housing supply ratio is affected by unemployment, we focus on the lower left graph in Plot 24. We see that the first significant spike occurs at lag 1 (excluding lag 0). In addition, that significant spike decays right away, so we know housing supply ratio at time t depends on unemployment at time (t-1).

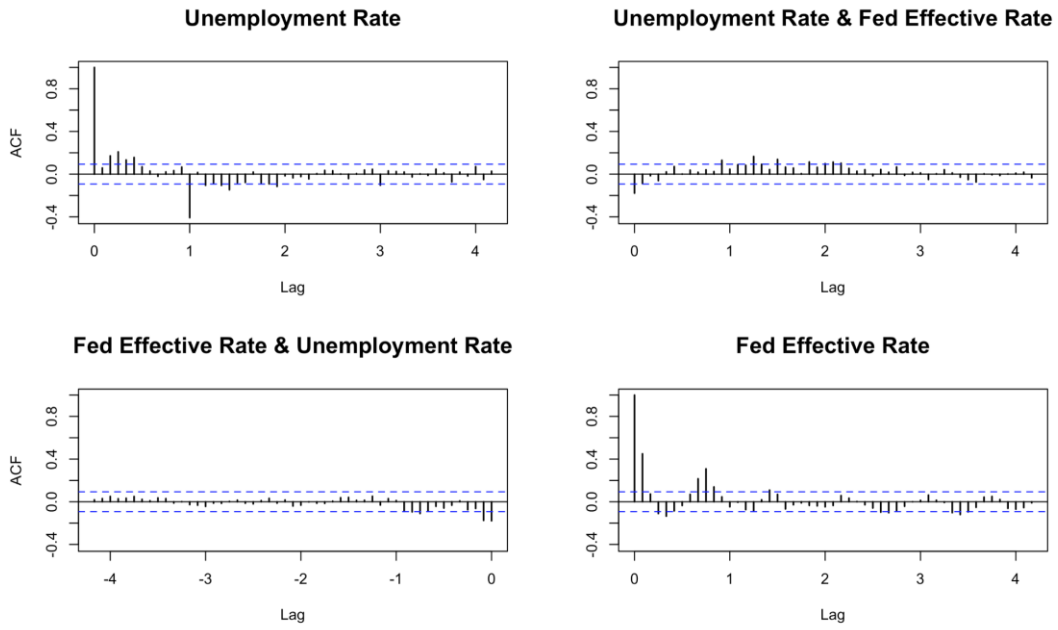
Next, we see how the housing supply ratio is related to itself. We see that CCFs die away in a damped sine-wave fashion, so we know housing supply ratio at time t depends on housing supply ratio at time t-1 and housing supply ratio at time t-2.

Writing this in formula form, we have:

$x^*_{1,t} = a_1 y^*_{t-1} + a_2 x^*_{1,t-1} + a_3 x^*_{1,t-2}$.

That is to say, we fit a VAR(6) model between unemployment and housing supply ratio with:

$x^*_{1,t} = a_1 y^*_{t-1} + a_2 x^*_{1,t-1} + a_3 x^*_{1,t-2}$ and $y^*_t = a_1 x^*_{1,t-6} + a_2 y^*_{t-1} + a_3 y^*_{t-2}$



Plot 25: CCF between Unemployment and Federal Reserve Effective Rate training data

From Plot 25, we see the cross-correlation between unemployment and federal reserve effective rate. Based on this, we can find the VAR model between them. Since we care about how the federal reserve effective rate affects unemployment, we look at the top right graph in Plot 25. We see that the first significant spike occurs at lag 1 (excluding lag 0).

In addition, that significant spike decays right away, so we know unemployment at time t dependents on federal reserve effective rate at time (t-1). Next, we see how unemployment is related to itself. We see that CCFs die away in a damped sine-wave fashion, so we know unemployment at time t depends on unemployment at time t-1 and unemployment at time t-2.

Writing this in formula form, we have:

$$y^*_t = a_1 x^*_{2,t-1} + a_2 y^*_{t-1} + a_3 y^*_{t-2}$$

When we look at how the federal reserve rate is affected by unemployment, we focus on the lower left graph in Plot 25. We see that the first significant spike occurs at lag 1 (excluding lag 0). In addition, that significant spike decays right away, so we know the reserve rate at time t depends on unemployment at time (t-1).

Next, we see how the federal reserve rate is related to itself. We see that CCFs die away in a damped sine-wave fashion, so we know federal reserve rate at time t depends on federal reserve rate at time t-1 and federal reserve rate at time t-2.

Writing this in formula form, we have:

$x^*_{2,t} = a_1 y^*_{t-1} + a_2 x^*_{2,t-1} + a_3 x^*_{2,t-2}$.

[Here $x^*_{2,t}$ refers to differenced federal reserve rate training data]

This means that we need to fit a VAR(2) model between unemployment and federal reserve effective rate. By looking at both Plot 24 and Plot 25, we can see that unemployment is leading since it is significant at lag 1 in the lower left graph, while in the upper right graph, housing is not significant until lag 6.

In Plot 25, we see that in both the upper right and lower left graph, lag 1 is significant, but the lower left graph has a more significant spike than the upper right graph, meaning unemployment is more "leading" than federal reserve rate.

IX.2 VAR Model

In this section, we fit the VAR(6) and VAR(2) models:

1. **VAR(6) Model**

Unemployment:

	Estimate		**Estimate**
Unemployment.Rate.l1	-0.057276934	Unemployment.Rate.l4	0.149502297
Housing.Supply.Raio.l1	0.036779918	Housing.Supply.Ratio.l4	0.028874474
Unemployment.Rate.l2	0.138770027	Unemployment.Rate.l5	0.143440787
Housing.Supply.Ratio.l2	0.060761740	Housing.Supply.Ratio.l5	0.035424376
Unemployment.Rate.l3	0.222121689	Unemployment.Rate.l6	0.052098039
Housing.Supply.Ratio.l3	0.040446479	Housing.Supply.Ratio.l6	0.044845204
const	0.001448732		

Housing Supply Ratio:

	Estimate		**Estimate**
Unemployment.Rate.l1	-0.628678932	Unemployment.Rate.l4	0.061052072
Housing.Supply.Ratio.l1	-0.247767338	Housing.Supply.Ratio.l4	-0.153412526
Unemployment.Rate.l2	-0.478534543	Unemployment.Rate.l5	-0.121499075

Housing.Supply.Ratio.l2	-0.188153630	Housing.Supply.Ratio.l5	-0.048113699
Unemployment.Rate.l3	0.058333071	Unemployment.Rate.l6	-0.534840301
Housing.Supply.Ratio.l3	0.045247464	Housing.Supply.Ratio.l6	-0.066892020
const	-0.008287217		

With the information about the model coefficients, we conclude the system is:

$$x^*_{1,t} = -0.2478 * x^*_{1,t-1} - 0.1882 * x^*_{1,t-2} + 0.04524 * x^*_{1,t-3} - 0.1534 * x^*_{1,t-4} - 0.04811 * x^*_{1,t-5} - 0.0669 * x^*_{1,t-6} - 0.6287 * y^*_{t-1} - 0.4785 * y^*_{t-2} + 0.05833 * y^*_{t-3} + 0.06105 * y^*_{t-4} - 0.1215 * y^*_{t-5} - 0.5348 * y^*_{t-6} - 0.00829$$

$$y^*_{t} = -0.05728 * y^*_{t-1} + 0.13877 * y^*_{t-2} + 0.2221 * y^*_{t-3} + 0.1495 * y^*_{t-4} + 0.14344 * y^*_{t-5} + 0.05210 * y^*_{t-6} + 0.03677 * x^*_{1,t-1} + 0.06076 * x^*_{1,t-2} + 0.04044 * x^*_{1,t-3} + 0.02887 * x^*_{1,t-4} + 0.04811 * x^*_{1,t-5} + 0.04485 * x^*_{1,t-6} + 0.001449$$

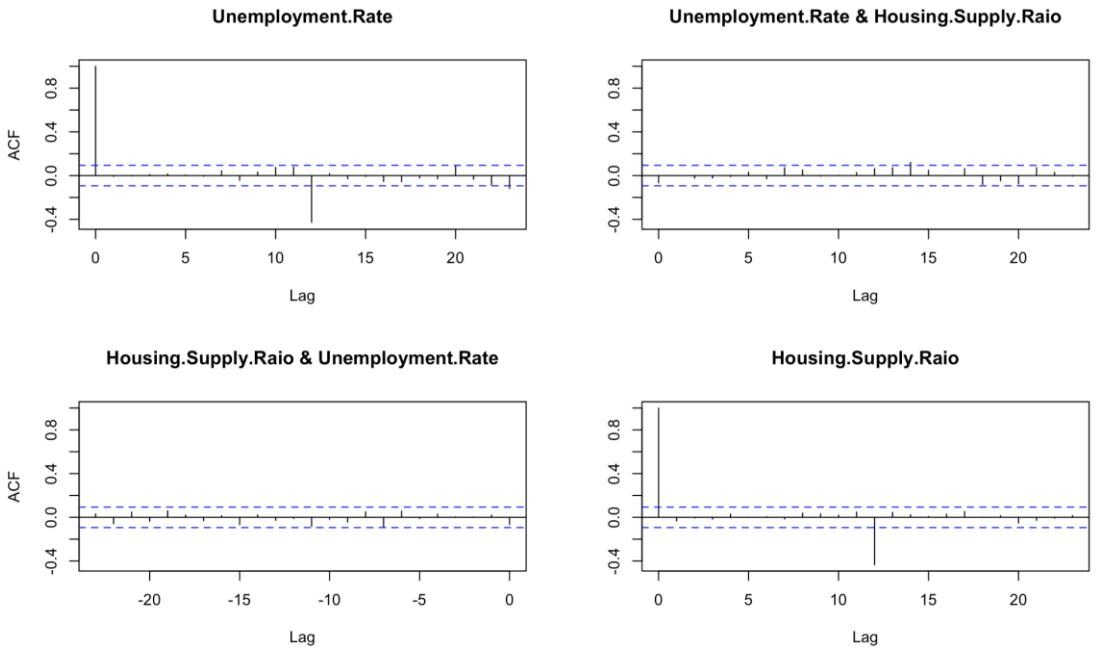
2. VAR(2) Model

Unemployment Rate: *Federal Funds Effective Rate:*

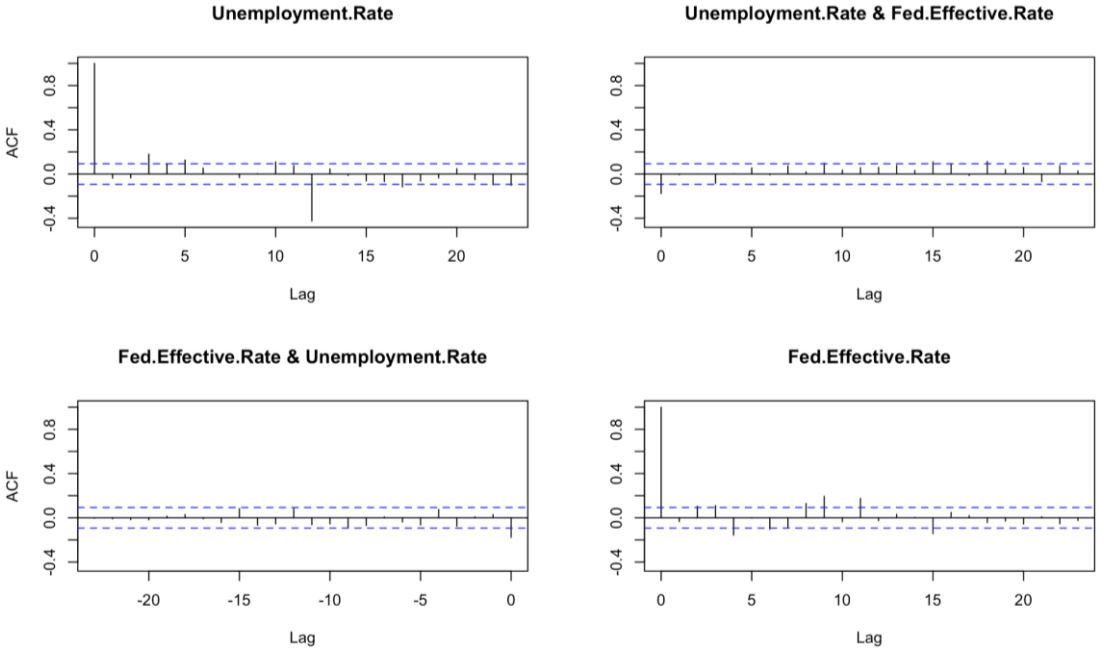
	Estimate		**Estimate**
Unemployment.Rate.l1	0.034332129	Unemployment.Rate.l1	-0.146505553
Fed.Effective.Rate.l1	-0.068430632	Fed.Effective.Rate.l1	0.535638790
Unemployment.Rate.l2	0.162287671	Unemployment.Rate.l2	0.007378826
Fed.Effective.Rate.l2	0.041075519	Fed.Effective.Rate.l2	-0.175693686
const	-0.003645138	const	-0.020028198

With the information about the model coefficients, we conclude the system is:

$$x_{2,t} = 0.5356 * x_{2,t-1} - 0.1757 * x_{2,t-2} - 0.1465 * y_{t-1} + 0.00738 * y_{t-2} - 0.02003$$
$$y_t = 0.0343 * y_{t-1} + 0.1623 * y_{t-2} - 0.0684 * x_{2,t-1} + 0.0411 * x_{2,t-2} - 0.003645$$



Plot 26: ACF of Residuals for VAR(6)

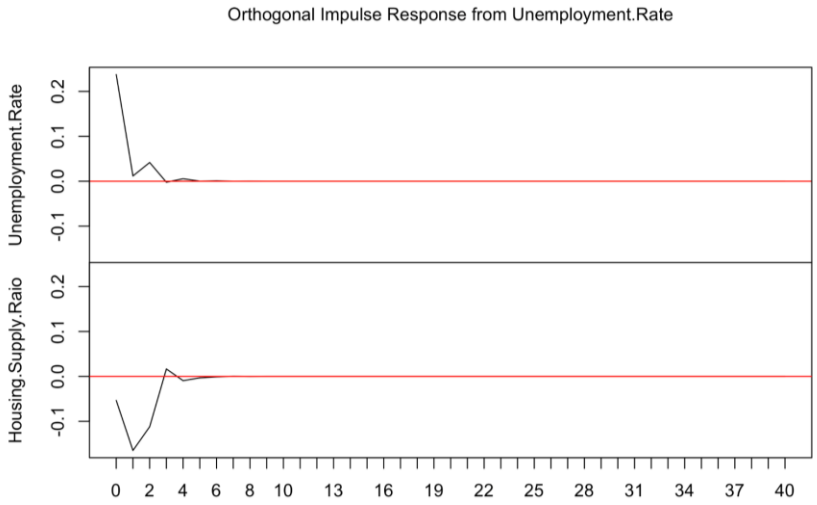


Plot 27: ACF of Residuals for VAR(2)

By looking at the residuals of both our VAR(6) and VAR(2) models in Plot 26 and Plot 27, since the residuals are almost bivariate white noise, the assumption for the VAR model is validated.

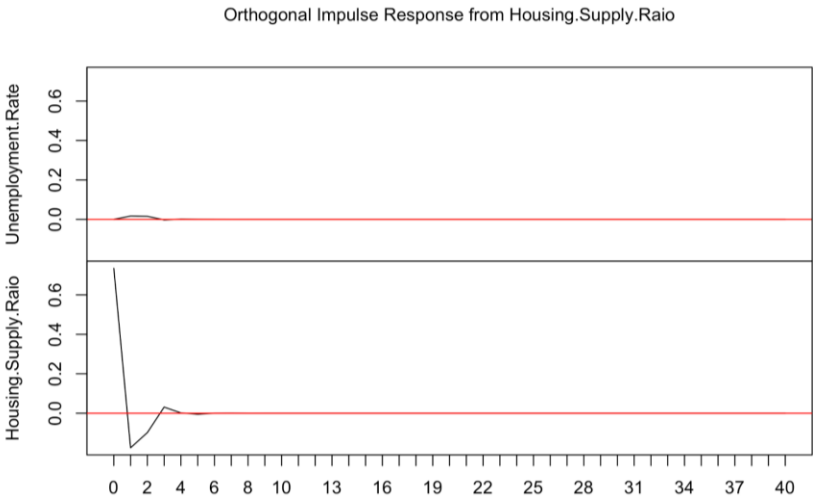
IX.3 Impulse Response Functions

In this section, we construct and plot the impulse response functions to see the effect and length of the effect of a shock to the system.



Plot 28: Orthogonal Impulse Response for Unemployment Rate

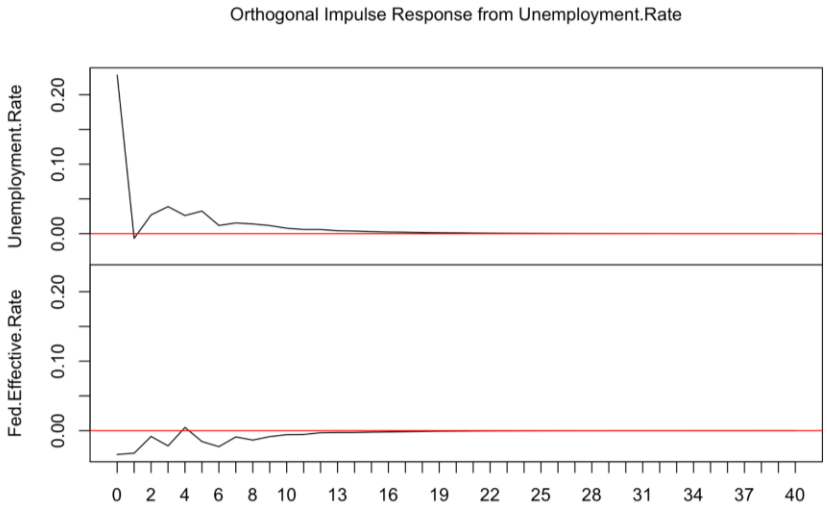
In Plot 28, we can see how Unemployment rate and Housing Supply Ratio respond to an impulse in Unemployment rate. Notice both of them have a relatively large response, and the equilibrium is achieved after t = 4.



Plot 29: Orthogonal Impulse Response for Housing Supply Ratio

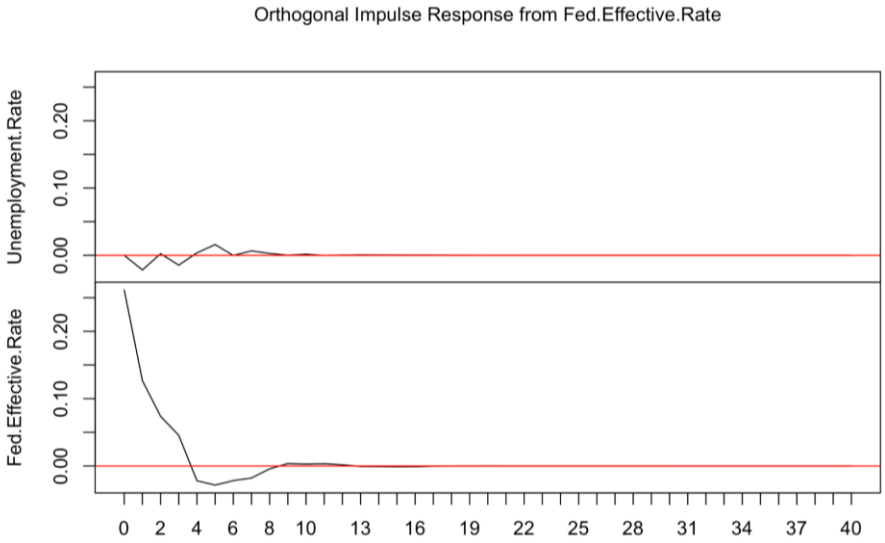
In Plot 29, we can see how the Unemployment Rate and Housing Supply Ratio respond to an impulse in Housing Supply Ratio. Unemployment does not respond much to the shock, while the Housing Supply Ratio responds dramatically to the impulse.

Therefore, Equilibrium is achieved after t = 4.



Plot 30: Orthogonal Impulse Response from Unemployment Rate

In Plot 30, we can see how Unemployment Rate and Federal Reserve Effective Rate respond to an impulse in Unemployment rate. Clearly, the Unemployment Rate responds drastically to the impulse, but the Federal Reserve Effective Rate responds more mildly.
In conclusion, both of them reach equilibrium at t = 10.

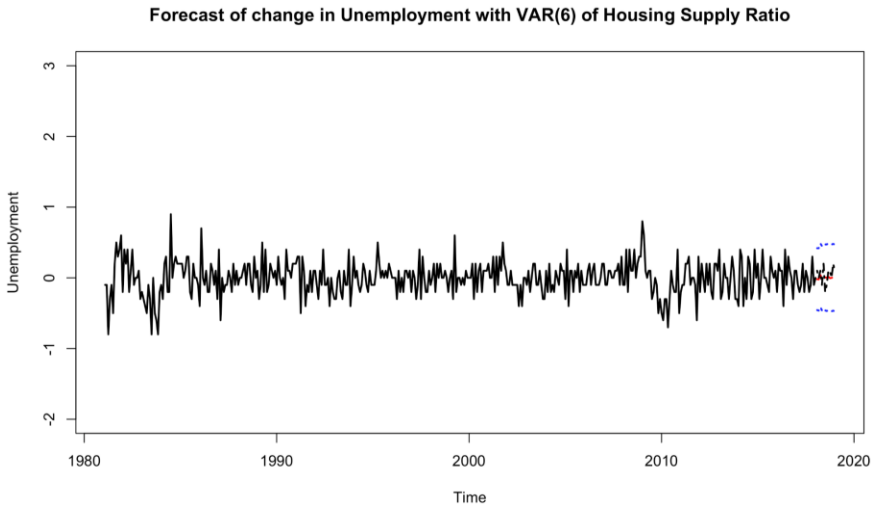


Plot 31: Orthogonal Impulse Response for Federal Funds Rate

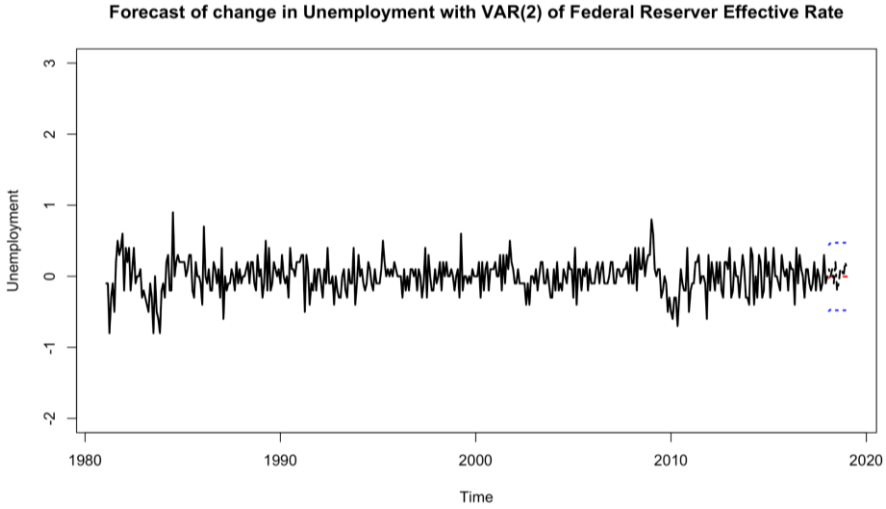
Finally, in Plot 31, we can see how the Unemployment rate and Federal Reserve Effective Rate respond to an impulse in Federal Reserve Effective Rate. The Unemployment Rate does not respond much to the impulse, but the Federal Reserve Effective Rate responds quite dramatically. Both of them reach equilibrium at t = 10.

IX.4 Forecasting

After figuring out the VAR models, we can use them to make predictions. In particular, we are interested in how well the VAR(6) and VAR(2) models can predict the unemployment rate.In Plot 32 and Plot 33, we draw the prediction together with the confidence interval for the prediction for both VAR(6) and VAR(2).



Plot 32: Forecast of Change in Unemployment with VAR(6) of Housing Supply Ratio

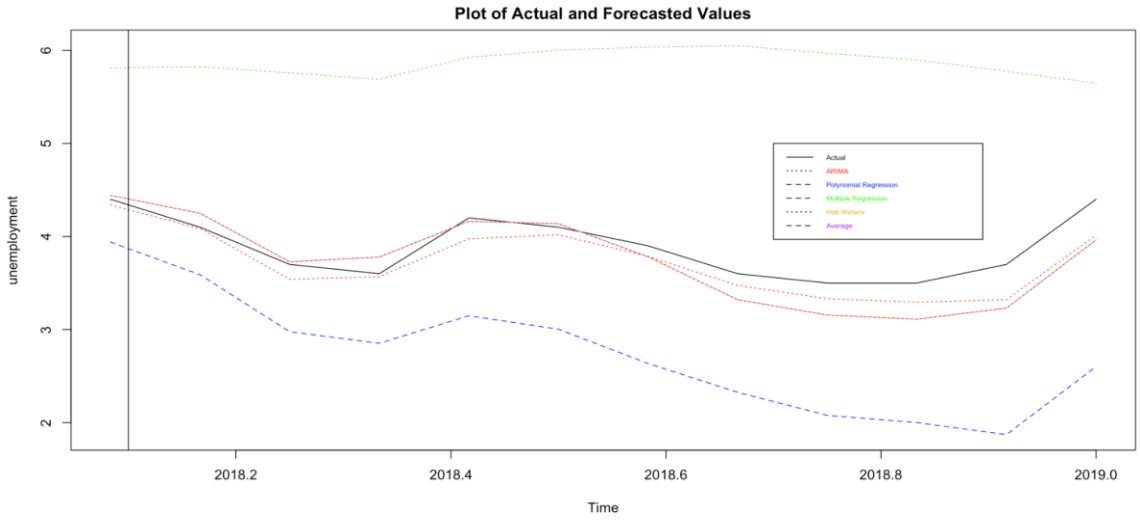


Plot 33: Forecast of Change in Unemployment with VAR(2) of Federal Reserve Rate

If we calculate the RMSE, we have the RMSE for the VAR(6) model being **0.4184812**, and the RMSE for the VAR(2) model being **0.4452308**.

By comparing the RMSE, we can see the VAR(6) model predicting Unemployment with Housing Supply Ratio is a better model than the VAR(2) model predicting Unemployment with Federal Reserve Effective Rate.

X. Forecast Comparison, Final Conclusion



Plot 34: Plot of Actual and all Forecasted Values

The RMSE for the exponential smoothing model is **0.2632173**, the RMSE for the polynomial regression model is **1.222444**, the RMSE for the GLS model is **2.328397,** the RMSE for the ARIMA model is **0.5637**, the RMSE for the VAR model is **0.41848,** and the average forecast RMSE is **0.263217** (Table 7). We can say that the exponential smoothing model performed the best so far in fitting the unemployment rate data and the multiple linear regression model (GLS) the worst so far.

Notice that we did not include the VAR models in Plot 34. The reason is that we have applied different differencing methods to our dependent variable and independent variables to make them stationary before fitting the VAR models. That is why Plot 32 and Plot 33 look different from the raw time plot in Plot 2 even though they are all time plots for unemployment rate. The only difference is that Plot 32 and Plot 33 show the differenced unemployment rate. That is to say, the testing data for dependent variables and independent variables have been differenced as well.

One should notice that the **Diff Raw Data column** in table 7 are completely different from the **Raw Data Values** since the former one is the differenced version of the latter one. That's why we choose not to plot VAR in Plot 34. If we really want to do so, we have to undo the differencing, which is quite laborious. In table 7, we have attached the VAR prediction from our VAR(6) since it has a smaller RMSE than the VAR(2) model.

Date	Raw Data Values	Exponential Smoothing	Polynomial Regression	Multiple Regression	Diff. Raw Data	VAR	ARIMA	Average Forecast
2018, 2	4.4	4.441674	3.939979	6.247967	0.1	-0.01856	4.69013	4.441674
2018, 3	4.1	4.249191	3.589400	6.278818	0	-0.0221	4.58632	4.249191
2018, 4	3.7	3.728389	2.977189	6.263476	0.1	0.03820	4.18789	3.728389
2018, 5	3.6	3.779068	2.851970	6.263235	-0.1	-0.0056	4.37285	3.779068
2018, 6	4.2	4.163007	3.148521	6.234664	0.2	0.0274	4.93323	4.163007
2018, 7	4.1	4.137926	3.003566	6.210484	-0.2	0.00535	5.12979	4.137926
2018, 8	3.9	3.781335	2.635782	6.184804	-0.1	0.00179	5.05041	3.781335
2018, 9	3.6	3.320676	2.325321	6.172834	0.1	0.01552	4.88785	3.320676
2018, 10	3.5	3.155865	2.078094	6.141373	0.1	-0.0047	4.8918	3.155865
2018, 11	3.5	3.111732	1.999822	6.144801	0.0	-0.0007	5.00000	3.111732
2018, 12	3.7	3.230751	1.871225	6.091753	0.2	0.00705	5.16746	3.230751
2019, 1	4.4	3.963826	2.602641	6.161996	0.1	-0.0011	5.99869	3.963826
	RMSE	**0.263217**	**1.222444**	**2.328397**		**0.41848**	**0.5637**	**0.263217**

Table 7: Actual Values, Forecasted Values, and respective RMSE