STA137 Project II

Junyao Lu

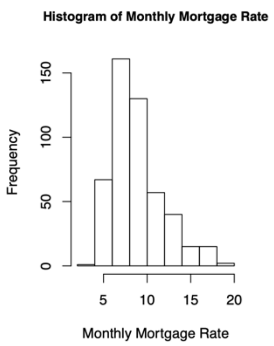
915938649

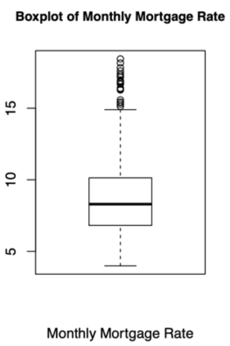
Introduction

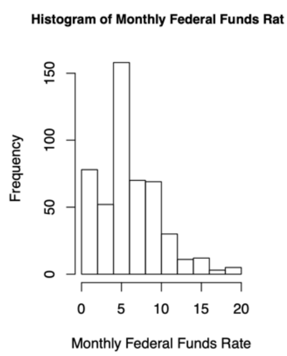
A mortgage is known as a type of loan that people can borrow money from banks and financial institutions to buy or refinance a home or piece of property. The mortgage rate is the interest rate of the mortgage, which can be fixed or fluctuate depending on the agreements. In order to understand the past mortgage rates, predict future mortgage rates, and make relative policy suggestions, the dataset of the US monthly 30-year conventional mortgage rates from April 1971 to November 2011 is obtained from Federal Reserve Economic Data to develop a precise statistical model. I am also interested in investigating the relationship between the mortgage rates and the Federal Funds rates. There are in total of 488 observations in the dataset collected from 1971 to 2011. There are 5 variables in the dataset, which are year, month, day, morg known as the monthly mortgage rate, and ffr known as the monthly federal funds rate. There is one additional date variable constructed based on the values of variables year, month, and day, which is useful for the time series analysis. In the following report, I am going to examine the stationarity of the data, transform the data if it is not stationary, build ARIMA models, perform model checking and selection, and fit a time series model for the mortgage rate and the lag-1 federal funds.

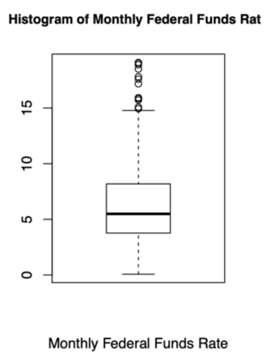
Material and Methods

This dataset is time-series data because each observation is collected at successive time periods, which is the first day of each month from April 1971 to November 2011. For the variable morg, the minimum monthly mortgage rate is 3.990, the maximum monthly mortgage rate is 18.450, and the mean is 8.800. From the histogram, I can see that the monthly mortgage rate is right skewed, which means that there are a lot of observations that are larger than the mode. The boxplot reveals the same information that there are a lot of outliers beyond the upper whisker. For the variable ffr, the minimum monthly federal funds rate is 0.070, the maximum monthly federal funds rate is 19.100, and the mean is 5.995. The histogram of the monthly federal funds rate also shows a right skewness, and the boxplot points out a lot of outliers beyond the upper whisker.

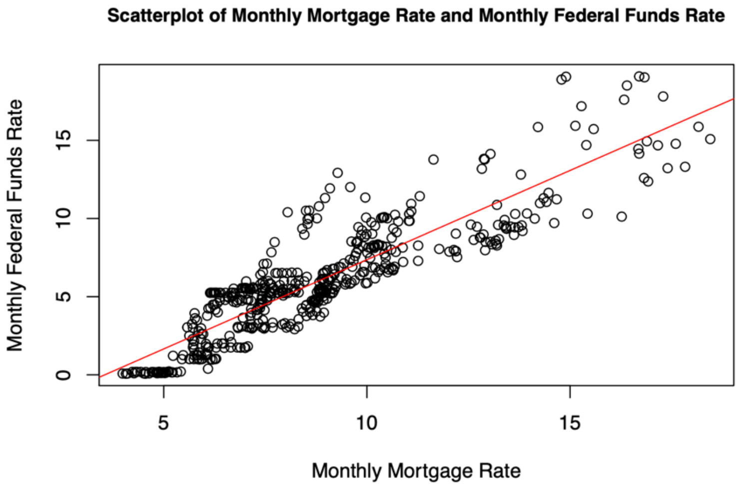


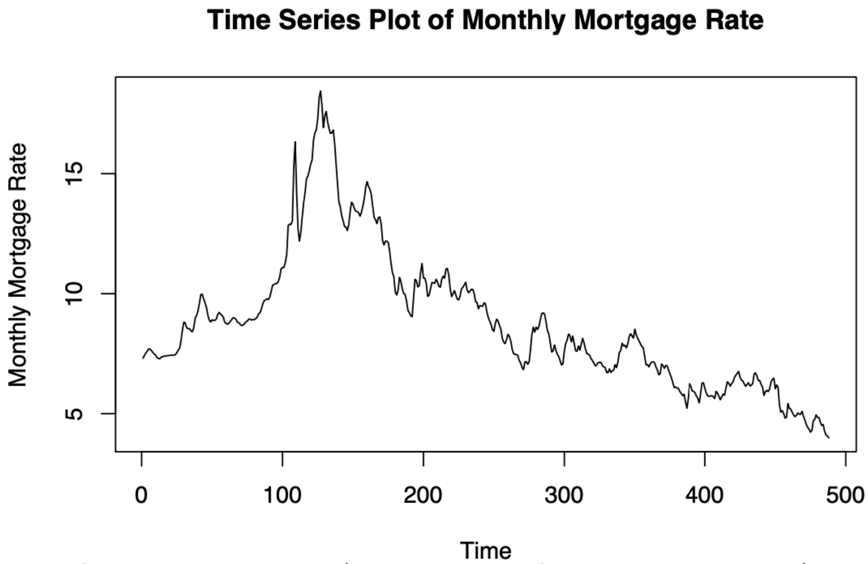


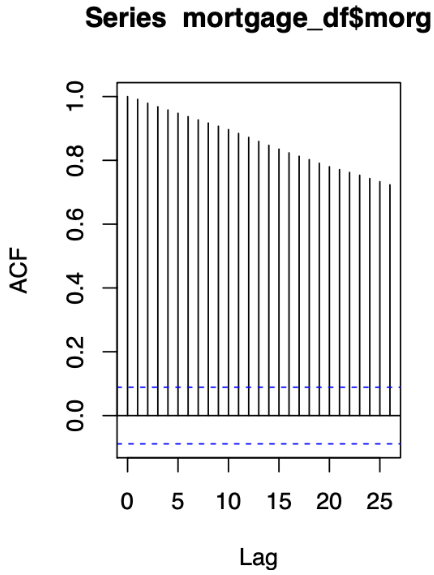


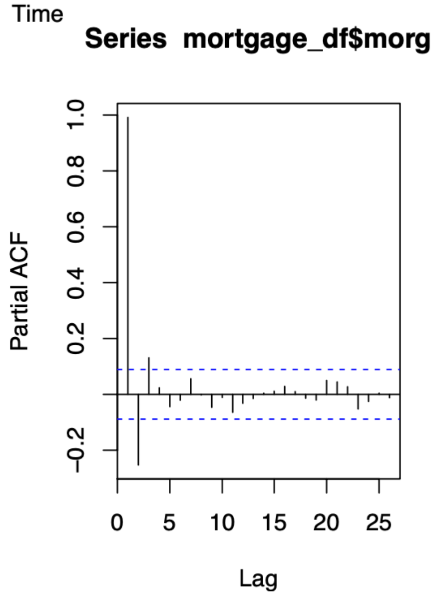


The scatter plot of these two variables shows that there is a positive linear relationship between the monthly mortgage interest rate and the monthly federal funds rate. As the monthly mortgage interest rate increases, the monthly federal funds rate also tends to increase.





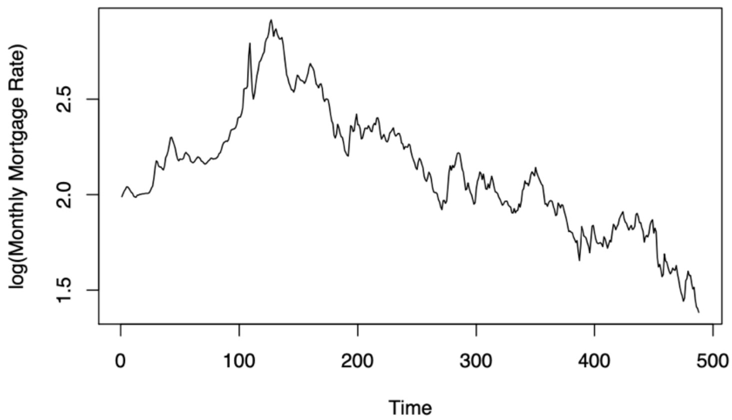




Then, from the time series plot of the monthly mortgage interest rate, it can be seen that there is a clear trend in the data, and there is no constant mean and variance. So, the data is not stationary. The ACF ordinates are large and decay slowly, and the PACF cuts off after lag 1. These observations also confirm that the data is not stationary.
In order to fix the problem of non-stationarity, I can apply transformations to the original data, such as log transformation, log differencing, and differencing. Next, I am going to apply each of these transformations to the data to see if they can make the transformed series stationary.

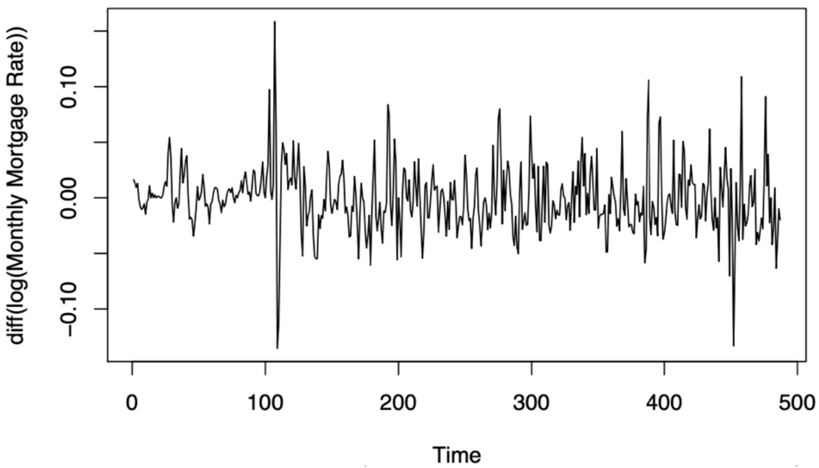
Log Transformation

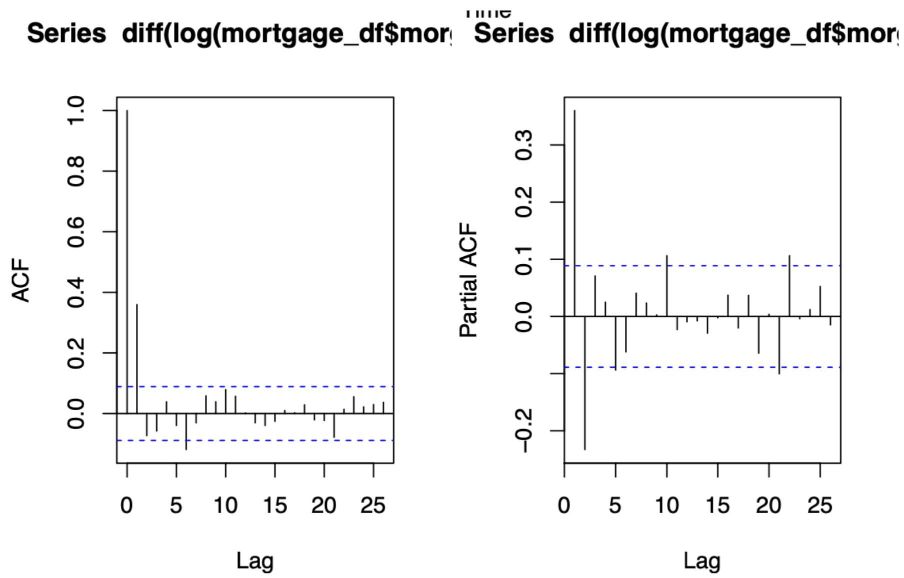
After applying log transformation to the data, I can see that there is still an obvious trend in the plot and there is no constant mean and variance. The problem of non-stationarity is not fix by using log transformation. So, I will not continue using this method.



Log Differencing

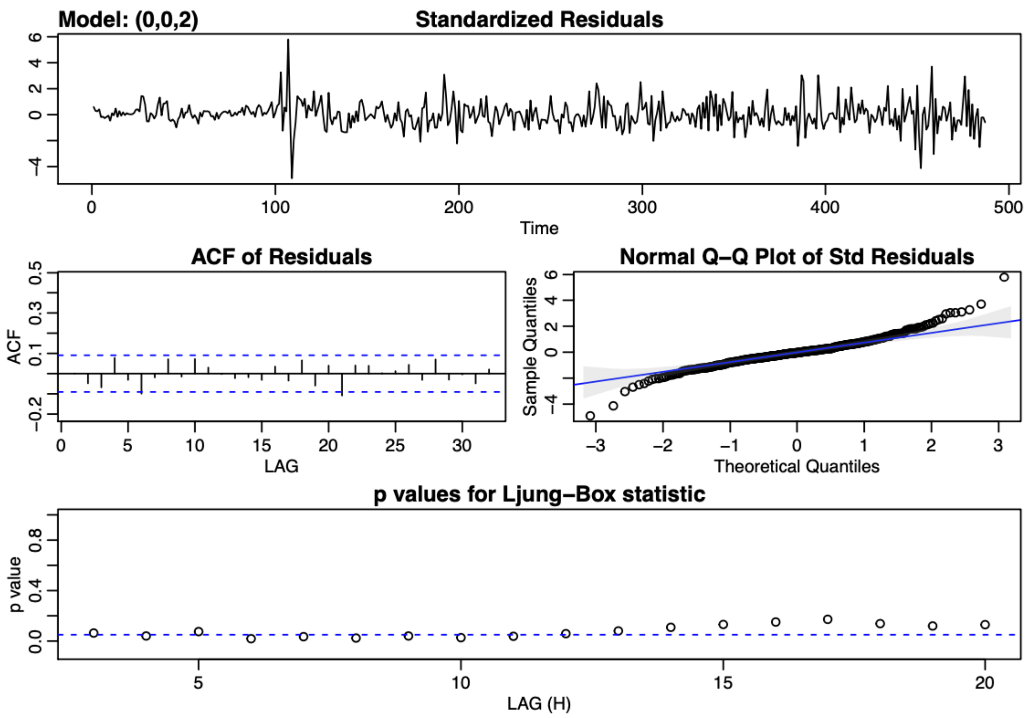
Then, I apply log differencing to the original data to see if this method helps. From the new time series plot, I observe that there is no significant trend, and the data is roughly stationary. The ACF and PACF plots of the log difference series suggest a MA(2) model and a ARMA(2, 2) model.





Then, I examine both models to see if they fit the data well.

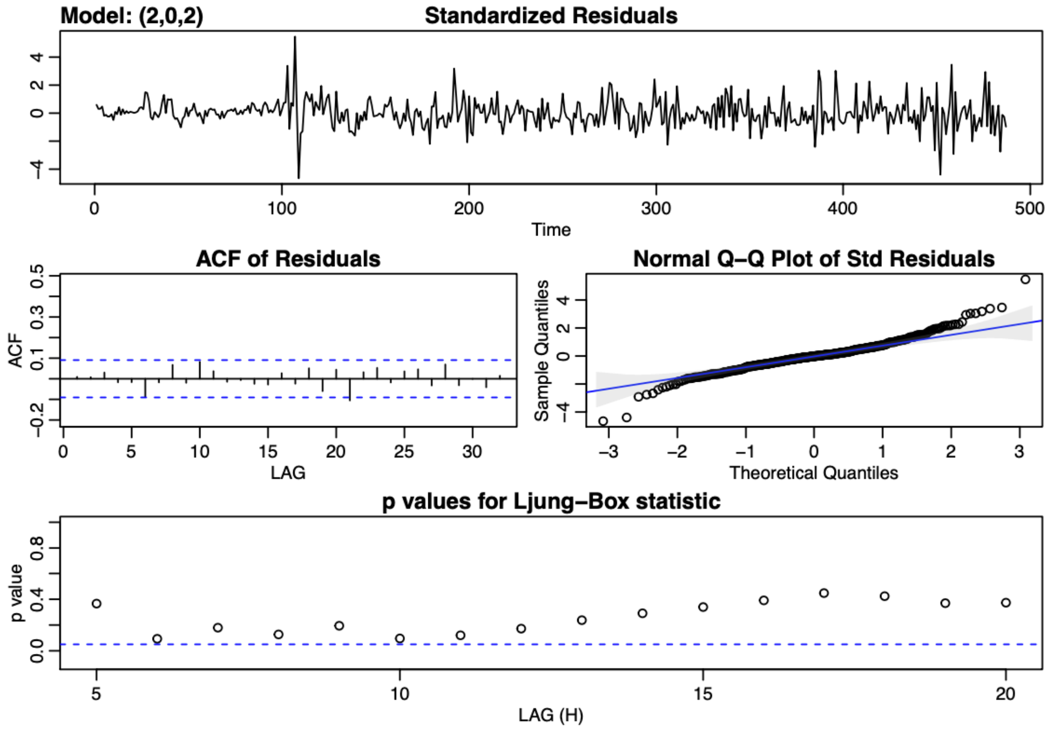
<u>Fitting MA(2) to the log difference series</u>



Estimated parameters of MA(2)

```
##        Estimate     SE t.value p.value
## ma1      0.4566 0.0484  9.4263  0.0000
## ma2      0.0008 0.0481  0.0166  0.9867
## xmean   -0.0012 0.0018 -0.6959  0.4868
```

Fitting ARMA(2, 2) to the log difference series



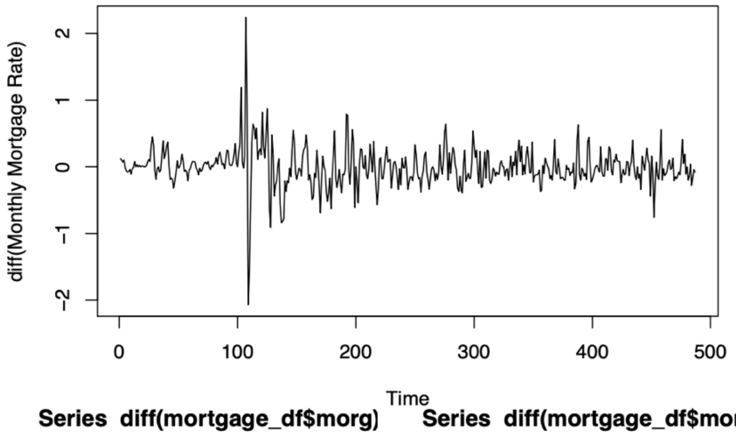
Estimated parameters of ARMA(2, 2)

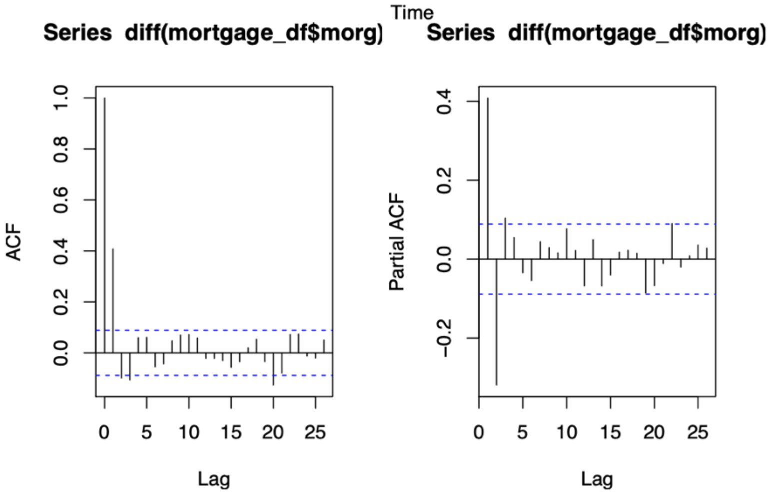
```
##       Estimate     SE t.value p.value
## ar1    -0.1598 0.1838 -0.8693  0.3851
## ar2    -0.3093 0.1149 -2.6920  0.0073
## ma1     0.6118 0.1799  3.4003  0.0007
## ma2     0.3345 0.1401  2.3871  0.0174
## xmean  -0.0013 0.0016 -0.7855  0.4326
```

From the results above, there is no apparent trend or pattern in both plots of the standardized residuals. For both models, the ACF shows no apparent significant dependence structure, as the ordinates are within the blue bounds. For the normal Q-Q plots, most of the points lie on the blue line. There are some deviations on both tails, probably due to the existence of outliers. But the normality assumption seems to be appropriate with the exception of outliers for both models. The only difference is that all the p-values for Ljung-Box statistics are above the blue dotted line for ARMA(2, 2), but some of the p-values for MA(2) are on or below the blue dotted line. Hence, ARMA(2, 2) for the log difference series is a better model for the data. From the estimated parameters for the model ARMA(2, 2), I can see that some of the p values are very large, which indicates that some of the estimated parameters are not statistically significant. So, this model may not be the best model to fit the data. I might need to consider refitting other models.

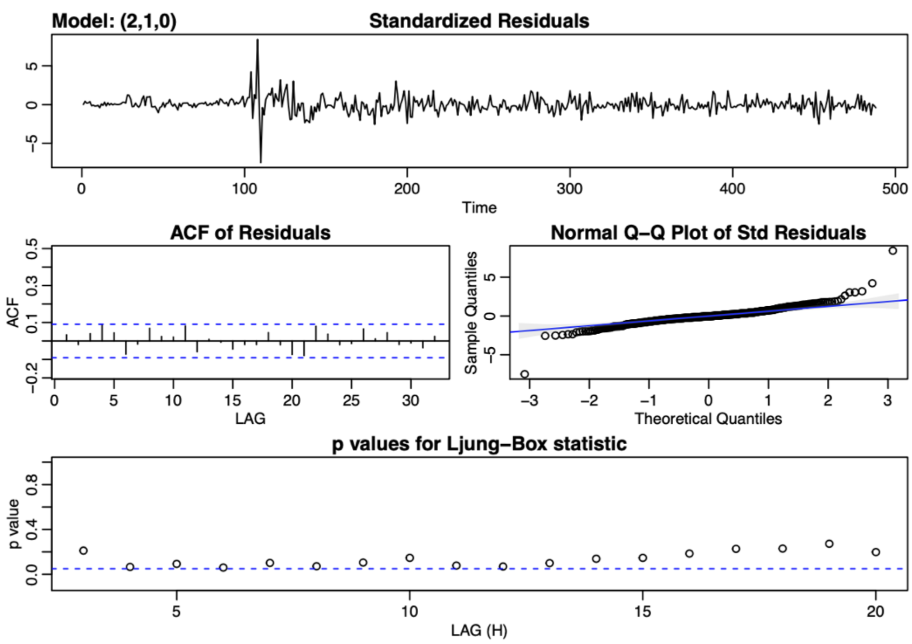
Differencing

Differencing is another transformation that can be used. By using the first-order difference, the data becomes roughly stationary in the new graph, and there is no obvious trend. The ACF and PACF plots suggest a ARIMA(2, 1, 0) model and a ARIMA(2, 1, 2) model to the original series.





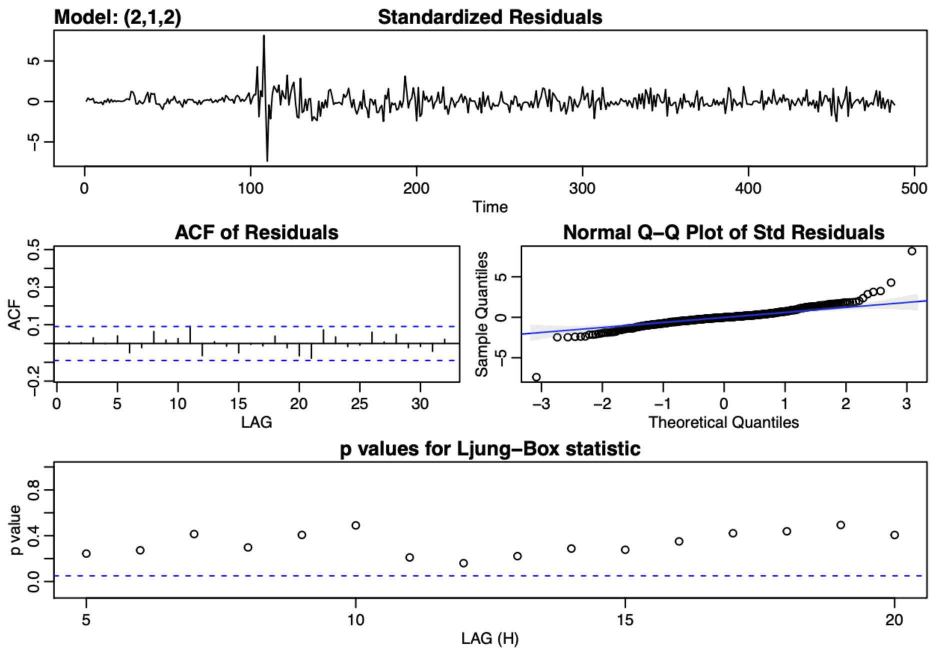
Fitting ARIMA(2, 1, 0)



Estimated parameters of ARIMA(2, 1, 0)

```
##           Estimate     SE t.value p.value
## ar1         0.5378 0.0429 12.5361   0.000
## ar2        -0.3182 0.0429 -7.4251   0.000
## constant   -0.0068 0.0149 -0.4610   0.645
```

Fitting ARIMA(2, 1, 2)



Estimated parameters of ARIMA(2, 1, 2)

```
##           Estimate     SE t.value p.value
## ar1         0.0860 0.1666  0.5162  0.6059
## ar2        -0.3260 0.0867 -3.7621  0.0002
## ma1         0.4736 0.1657  2.8575  0.0045
## ma2         0.2428 0.1240  1.9589  0.0507
## constant   -0.0068 0.0159 -0.4290  0.6681
```

From the results above, I conclude that there is no apparent trend or pattern in the plots of standardized residuals. For both models, the ACF shows no apparent significant dependence structure, as the ordinates are within the blue bounds. For the normal Q-Q plots, most of the points lie on the blue line. There are some deviations on both tails, probably due to the existence of outliers. But the normality assumption seems to be appropriate with the exception of outliers for both models. All the p-values for Ljung-Box statistics are above the blue dotted line, which indicates that the processes match the white noise process. Thus, both models fit well to the data and estimated parameters are significant.

In order to find the best fit model, I am going to perform model selection by comparing AIC, AICc, and BIC.

<u>ARIMA(2, 1, 0)</u>

AIC: 0.1279502

AICc: 0.1280522

BIC: 0.1623507

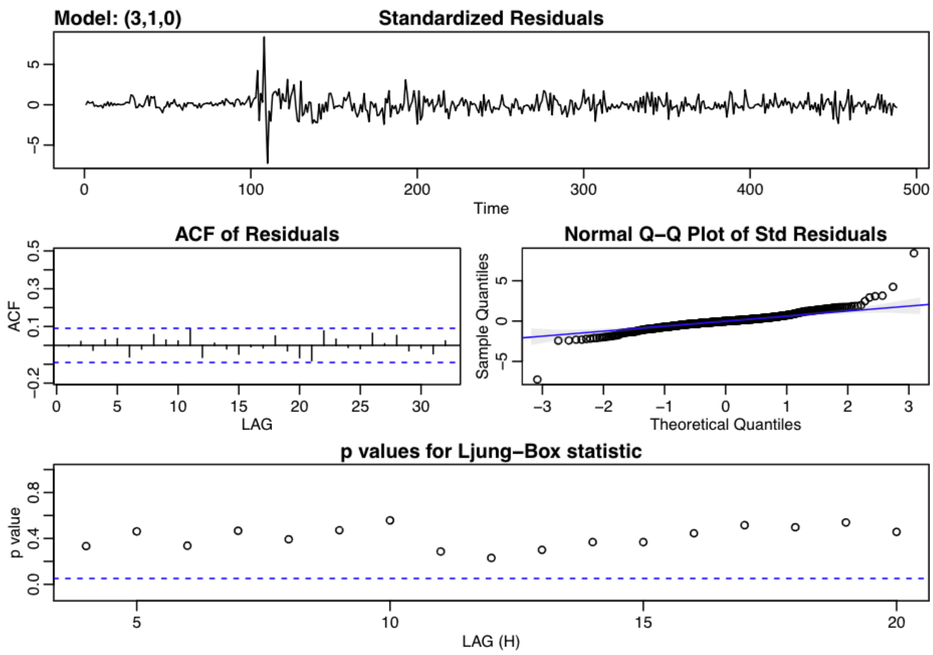
<u>ARIMA(2, 1, 2)</u>

AIC: 0.1229282

AICc: 0.1231843

BIC: 0.174529

From the above results, I notice that both AIC and AICc are smaller for ARIMA(2, 1, 2), which indicates that ARIMA(2, 1, 2) is a better fit. The BIC criterion is smaller for ARIMA(2, 1, 0). And the statistics are close to each other for both models, so they are both good fits. However, I also notice that some of the estimated parameters for ARIMA(2, 1, 2) are not significant. In this case, ARIMA(2, 1, 0) might be a better choice. As a final check, I might consider to overfit the model to see if the results change significantly. I find out that ARIMA(3, 1, 0) fits the data even better. Also, the AIC, AICc, and BIC are smaller for ARIMA(3, 1, 0) compared to others.



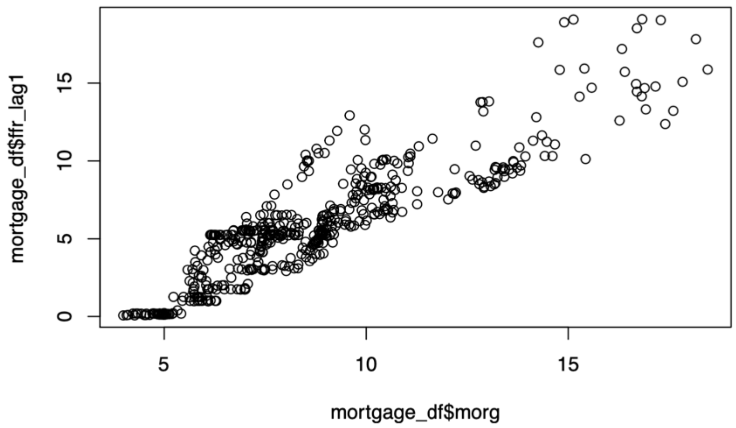

```
##           Estimate     SE t.value p.value
## ar1         0.5710 0.0450 12.6783  0.0000
## ar2        -0.3742 0.0490 -7.6293  0.0000
## ar3         0.1037 0.0450  2.3061  0.0215
## constant   -0.0068 0.0165 -0.4157  0.6778
```

AIC: 0.1212004; AICc: 0.1213708; BIC: 0.1642011

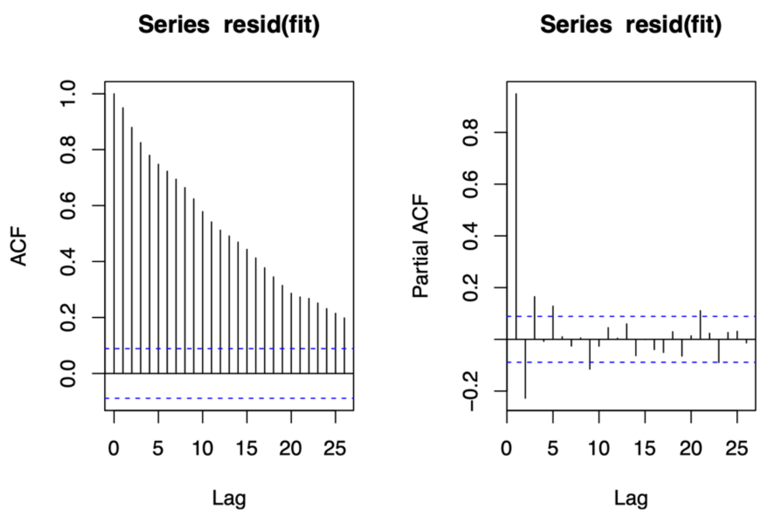
Therefore, I think ARIMA(3, 1, 0) is the best fit model.

$$x_t = 0.5710x_{t-1} + 0.5710x_{t-2} + 0.1037x_{t-3} + w_t$$

Next, I am going to find a time series model for the mortgage rate using the lag-1 federal funds rate as an explanatory variable. From the plot, I can see that there is a positive linear relationship between the mortgage rate and the lag-1 federal funds. This suggests that the model to consider could be $M_t = \beta_0 + \beta_1 F_{t-1} + x_t$



Then, based on the result, the plots of ACF and PACF are consistent with those of a AR(1) process.



Therefore, I am going to do the regression analysis with errors that follows a AR(1) process and I get the final fitted model is:

$$M_t = 4.62099 + 0.69606F_{t-1} + x_t$$

$$x_t = 0.9931x_{t-1} + w_t$$

where {wt} ~ WN(0, 0.07492)

Results

After conducting analysis to the dataset, I find out that data is not stationary. Then, I use different ways to transform the data in order to make it stationary, such as log transformation, log differencing, and differencing. The log transformation fails to adjust the stationarity of the data. The models built from the log differencing method have insignificant estimated parameters, so I refit the model using first-order differencing and find out ARIMA(3, 1, 0) is the best fit model for the data. This model fits the data well, has the lowest AIC, AICc and BIC, and has significant estimated parameters. The model can be written as $x_t = 0.5710x_{t-1} + 0.5710x_{t-2} + 0.1037x_{t-3} + w_t$. Then, I continue investigate the relationship between the mortgage rate and the lag-1 federal funds rate using regression with autocorrelated errors. The final fitted model I get is:

$$M_t = 4.62099 + 0.69606F_{t-1} + x_t$$

$$x_t = 0.9931x_{t-1} + w_t$$

where {wt} ~ WN(0, 0.07492)

Both of the models describe the data well and they can be used to understand the past mortgage rates, predict future mortgage rates, and make relative policy suggestions.

Appendix

```
mortgage_df = read.delim("mortgage.txt", sep = " ", header = TRUE)
mortgage_df$date = as.Date(ISOdate(year = mortgage_df$year,
                                   month = mortgage_df$month,
                                   day = mortgage_df$day))
dimension = dim(mortgage_df)
min = mortgage_df[1,1:2]
max = mortgage_df[488,1:2]
par(mfrow=c(1, 2))
hist(mortgage_df$morg, cex.main=0.9,
     xlab = "Monthly Mortgage Rate",
     main = "Histogram of Monthly Mortgage Rate")
boxplot(mortgage_df$morg, cex.main=0.9,
        xlab = "Monthly Mortgage Rate",
        main = "Boxplot of Monthly Mortgage Rate")
par(mfrow=c(1, 2))
hist(mortgage_df$ffr, cex.main=0.9,
     xlab = "Monthly Federal Funds Rate",
     main = "Histogram of Monthly Federal Funds Rate")
boxplot(mortgage_df$ffr, cex.main=0.9,
        xlab = "Monthly Federal Funds Rate",
        main = "Histogram of Monthly Federal Funds Rate")
par(mfrow=c(1, 1))
plot(mortgage_df$morg, mortgage_df$ffr, cex.main=0.9,
     xlab = "Monthly Mortgage Rate",
     ylab = "Monthly Federal Funds Rate",
     main = "Scatterplot of Monthly Mortgage Rate and Monthly Federal Funds Rate")
abline(lm(mortgage_df$ffr~mortgage_df$morg), col = "red")
par(mfrow=c(1, 1))
ts.plot(mortgage_df$morg, type = "l",
        ylab = "Monthly Mortgage Rate",
        main = "Time Series Plot of Monthly Mortgage Rate")
par(mfrow=c(1, 2))
acf(mortgage_df$morg)
pacf(mortgage_df$morg)
par(mfrow=c(1, 1))
ts.plot(log(mortgage_df$morg), type = "l", ylab = "log(Monthly Mortgage Rate)")
par(mfrow=c(1, 1))
ts.plot(diff(log(mortgage_df$morg)), type = "l",
        ylab = "diff(log(Monthly Mortgage Rate))")
par(mfrow=c(1, 2))
acf(diff(log(mortgage_df$morg)))
pacf(diff(log(mortgage_df$morg)))
fit_ma = astsa::sarima(diff(log(mortgage_df$morg)), 0, 0, 2)
fit_ma$ttable
fit_arma = astsa::sarima(diff(log(mortgage_df$morg)), 2, 0, 2)
fit_arma$ttable
par(mfrow=c(1, 1))
ts.plot(diff(mortgage_df$morg), type = "l",
        ylab = "diff(Monthly Mortgage Rate)")
par(mfrow=c(1, 2))
acf(diff(mortgage_df$morg))
pacf(diff(mortgage_df$morg))
fit_arima210 = astsa::sarima(mortgage_df$morg, 2, 1, 0)
fit_arima210$ttable
fit_arima212 = astsa::sarima(mortgage_df$morg, 2, 1, 2)
fit_arima212$ttable
fit_arima210$AIC
fit_arima210$AICc
fit_arima210$BIC
fit_arima212$AIC
fit_arima212$AICc
fit_arima212$BIC
fit_arima310 = astsa::sarima(mortgage_df$morg, 3, 1, 0)
fit_arima310
fit_arima310$ttable
fit_arima310$AIC
fit_arima310$AICc
fit_arima310$BIC
mortgage_df$ffr_lag1 = Hmisc::Lag(mortgage_df$ffr, 1)
plot(mortgage_df$morg, mortgage_df$ffr_lag1)
fit = lm(mortgage_df$morg ~ mortgage_df$ffr_lag1)
par(mfrow=c(1, 2))
acf(resid(fit))
pacf(resid(fit))
x = astsa::sarima(mortgage_df$morg, 1, 0, 0, xreg = mortgage_df$ffr_lag1)
```