Forecasting Time Series Project 2

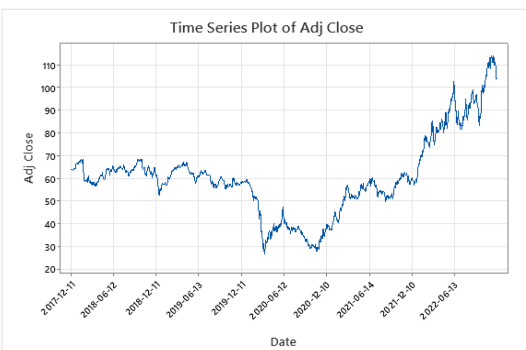
Yoon Tae Park(yp2201@nyu.edu)

Source of the data:
Exxon 5 year (17/12/11~22/12/08)
https://finance.yahoo.com/quote/Exxon/history?period1=1512950400&period2=1670716800&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

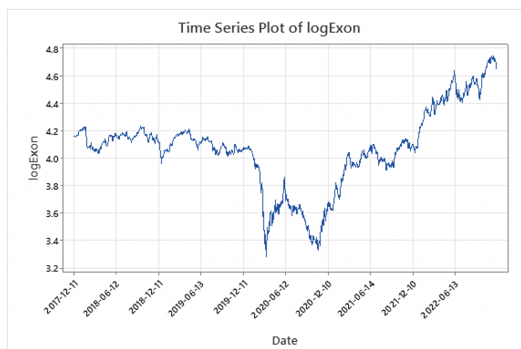
Q1. Find a time series data set which seems to show conditional volatility. The time series plot should show periods of relatively high volatility interspersed with periods of relative calm. It is often easier to identify conditional volatility if the sampling interval is short (e.g., daily data) than when the sampling interval is longer.

Ans)
I've selected the Exxon stock price for 5 years, and a time series data set seems to show conditional volatility. (Also for the log price time series) The graph indicates that the data fluctuates between periods of high volatility and periods of relative calm. Also, I will be using log stock price, as it shows less level-dependent volatility.

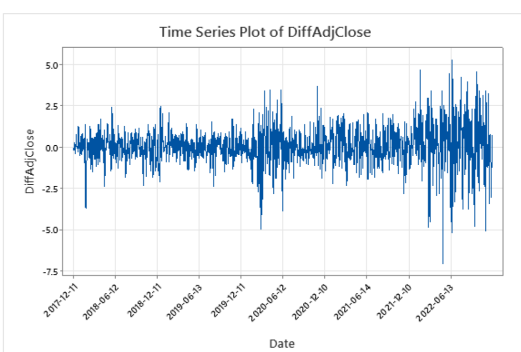
[Adj Close time series plot]



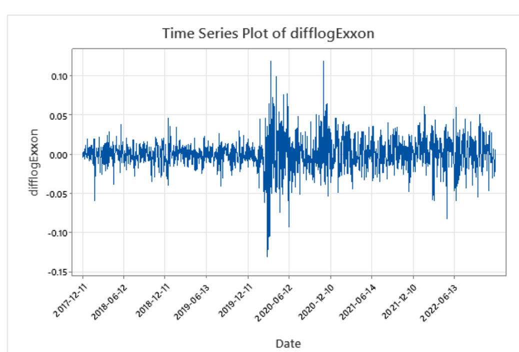
[Log Adj Close time series plot]



[Diff Adj Close time series plot]



[Diff Log Adj Close time series plot]

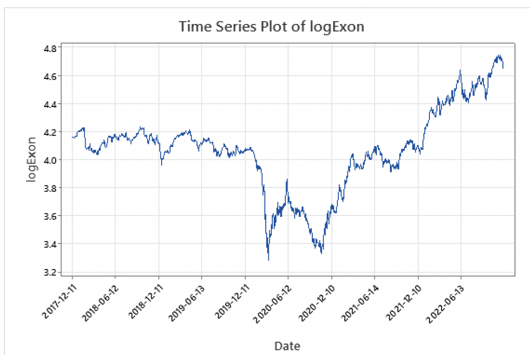


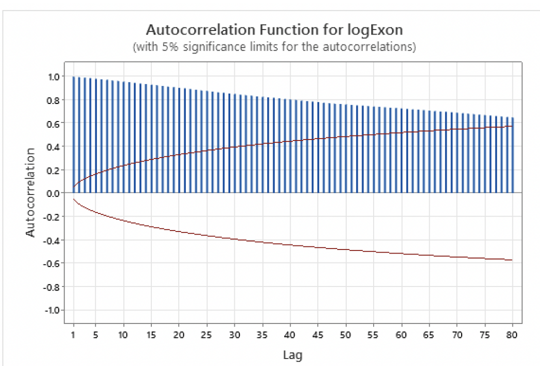
Q2. Leave out the last data point (let's call this the n +1'st observation) for the ARMAARCH modeling. You will use it at the end of the project for checking the performance of the forecast intervals. Using first n observations of your data set, carry out the same type of analysis as you did in Homework 8. Break your analysis into nine parts, just as in Homework 8

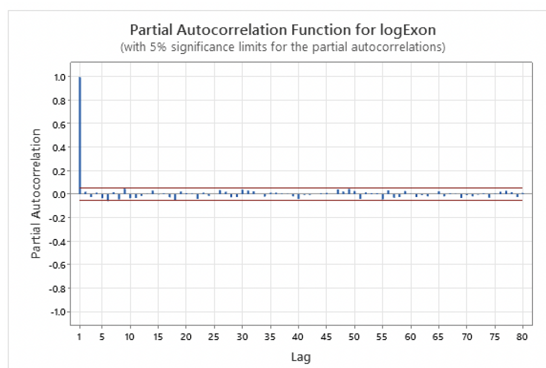
1) **Plot the logs of Exxon. Based on the plot, and the ACF and PACF of the logs and differenced logs, does the series appear to be stationary? Can you identify an ARIMA(p, d, q) model from these plots?**

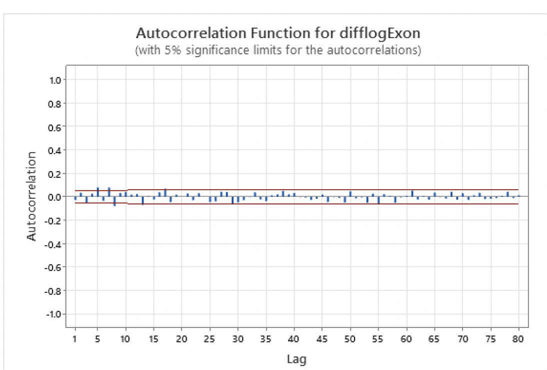
Ans)

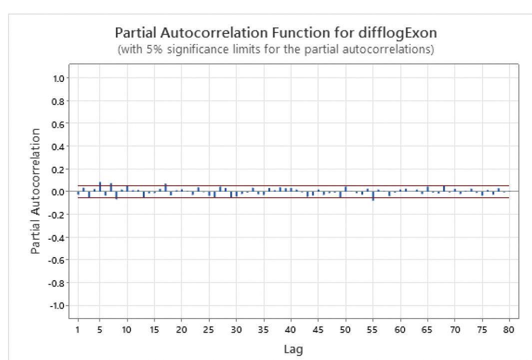
Based on the log Exxon time series plot, the series of log Exxon is not stationary. Based on ACF and PACF of log Exxon, we can see that PACF is showing a significant cut off after lag 1, and ACF of log Exxon dies down. Also, ACF and PACF of difference of log Exxon do not show any statistically significant lags. Also, we can see that the second difference of ACF and PACF is showing negative lags. Therefore, we can identify ARIMA (1,1,0).

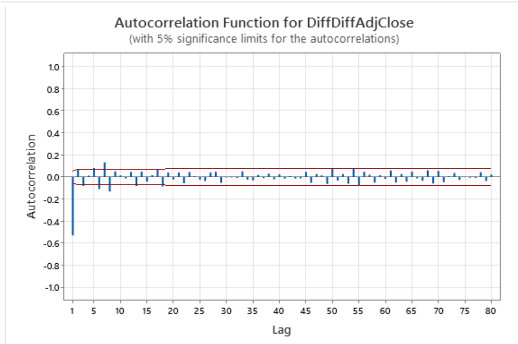


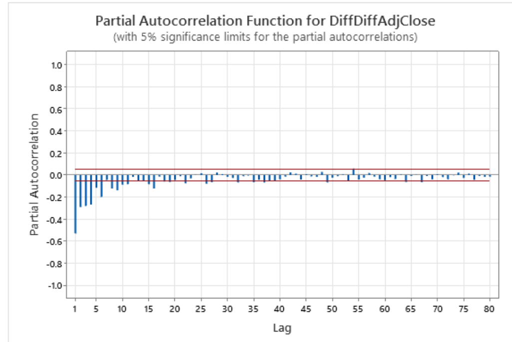












2) **Using AIcc, select an ARIMA(p, 1, q) with 0<=p<=2, 0<=q<=2. Write the complete form of the fitted model. Save the residuals and fitted values for the model you selected, using Storage -> residuals, Fits. The residuals will be stored in RESI1 and the fitted values will be stored in FITS1. (Note that FITS1 starts with one missing value, while at time t it represents f_t-1, 1, the one-step forecast for the log exchange rate at time t made from time t-1). Also, get Minitab to compute ARIMA one step ahead forecast, and 95% forecast interval.**

Ans)

Based on AICc scores, I've selected ARIMA (2,1,2) with constant as my best model. The complete form of the fitted model is as below:

$x_t = -1.4878x_{t-1} - 0.8575x_{t-2} + \varepsilon_t + 1.4567\varepsilon_{t-1} + 0.8679\varepsilon_{t-2} + 0.00131$, *where* x_t *= diff(log Exxon)*

[Best AICc, d=1, with constant]

Model Selection

Model (d = 1)	LogLikelihood	AICc	AIC	BIC
p = 2, q = 2*	3065.90	-6119.74	-6119.80	-6088.99
p = 0, q = 0	3055.51	-6107.01	-6107.02	-6096.74
p = 1, q = 0	3055.96	-6105.90	-6105.92	-6090.51
p = 0, q = 1	3055.93	-6105.84	-6105.86	-6090.45
p = 1, q = 1	3056.75	-6105.46	-6105.49	-6084.95
p = 2, q = 0	3056.68	-6105.32	-6105.36	-6084.81
p = 0, q = 2	3056.54	-6105.05	-6105.08	-6084.53
p = 1, q = 2	3057.45	-6104.86	-6104.90	-6079.22
p = 2, q = 1	3057.32	-6104.59	-6104.64	-6078.96

** Best model with minimum AICc. Output for the best model follows.*

Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	-1.4878	0.0426	-34.94	0.000
AR 2	-0.8575	0.0221	-38.84	0.000
MA 1	-1.4567	0.0371	-39.21	0.000
MA 2	-0.8679	0.0267	-32.56	0.000
Constant	0.00131	0.00197	0.67	0.504

Differencing: 1 Regular
Number of observations after differencing: 1257

Forecasts from Time Period 1258

			95% Limits		
Time Period	Forecast	SE Forecast	Lower	Upper	Actual
1259	4.65156	0.0211512	4.61010	4.69303	

Reference: [d=1, without constant] [d=0, with constant]

Model Selection

Model (d = 1)	LogLikelihood	AICc	AIC	BIC
p = 0, q = 0*	3055.30	-6108.59	-6108.59	-6103.46
p = 1, q = 0	3055.74	-6107.46	-6107.47	-6097.20
p = 0, q = 1	3055.71	-6107.41	-6107.42	-6097.14
p = 1, q = 1	3056.52	-6107.03	-6107.05	-6091.64
p = 2, q = 0	3056.47	-6106.92	-6106.94	-6091.53
p = 0, q = 2	3056.33	-6106.64	-6106.66	-6091.25
p = 2, q = 1	3057.11	-6106.19	-6106.22	-6085.68
p = 1, q = 2	3056.58	-6105.13	-6105.16	-6084.61
p = 2, q = 2	3057.30	-6104.55	-6104.60	-6078.92

* *Best model with minimum AICc. Output for the best model follows.*

Model Selection

Model (d = 0)	LogLikelihood	AICc	AIC	BIC
p = 1, q = 0*	3055.29	-6104.57	-6104.59	-6089.17
p = 2, q = 0	3055.69	-6103.35	-6103.38	-6082.83
p = 1, q = 1	3055.64	-6103.24	-6103.28	-6082.73
p = 2, q = 1	3056.49	-6102.92	-6102.97	-6077.28
p = 1, q = 2	3056.34	-6102.64	-6102.69	-6077.00
p = 2, q = 2	3057.26	-6102.45	-6102.51	-6071.69
p = 0, q = 2	1175.16	-2342.28	-2342.31	-2321.76
p = 0, q = 1	573.47	-1140.92	-1140.94	-1125.53

* *Best model with minimum AICc. Output for the best model follows.*

[d=0, without constant]

Model Selection

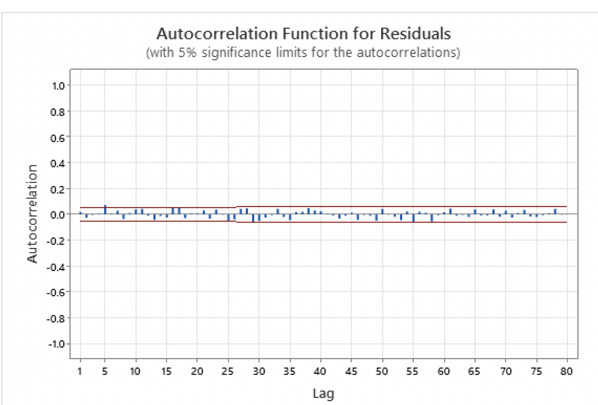
Model (d = 0)	LogLikelihood	AICc	AIC	BIC
p = 1, q = 0*	3048.16	-6092.32	-6092.33	-6082.05
p = 2, q = 0	2754.64	-5503.27	-5503.29	-5487.87

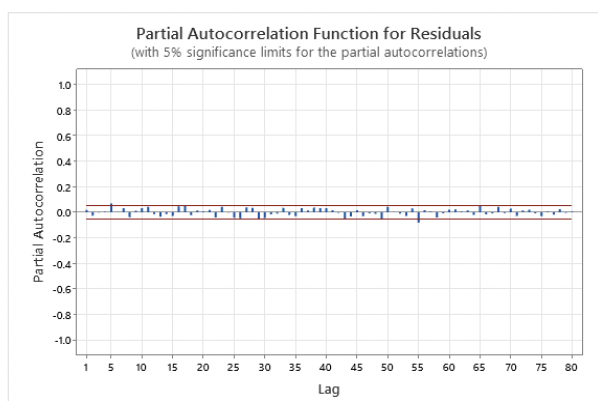
* *Best model with minimum AICc. Output for the best model follows.*

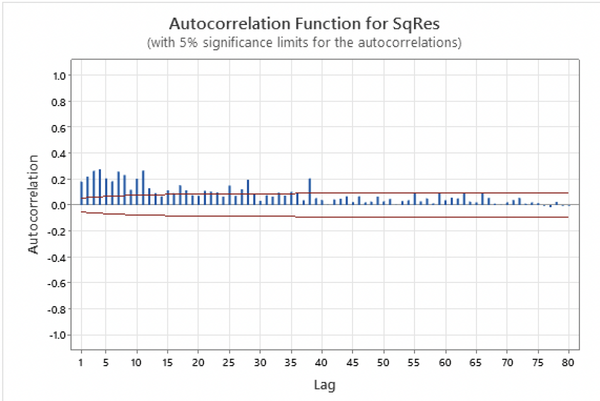
3) **Plot the residuals, as well as ACF and PACF of both the residuals and the squared residuals. Use these plots to argue that the residuals, although approximately uncorrelated, are not independent; instead, they show evidence of conditional heteroscedasticity.**

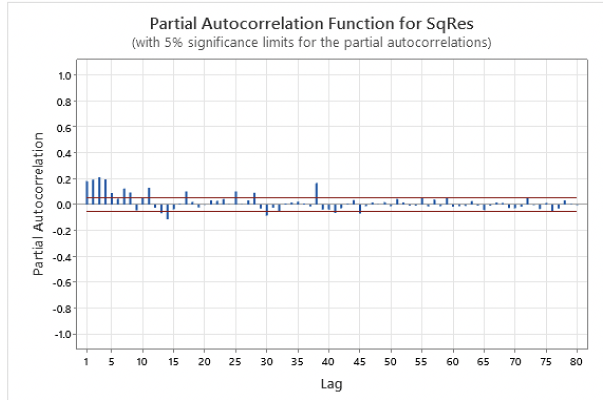
Ans)

ACF and PACF of Residuals for log Exxon don't have statistically significant lags, and seems like a white noise. However, ACF and PACF of the squared residuals shows multiple lags showing statistically significant. When we just look at the log Exxon, then they seem uncorrelated (as they seem like a white noise). However, when just squaring the residuals, they show some statistically significant points, showing evidence that they are not actually independent. It means that the variances are correlated, and this can be an evidence of conditional heteroscedasticity.









4) **Find the log likelihood values and AICc values for ARCH(q) models where q ranges from 0 to 10. Next, consider a GARCH(1,1) model. Evaluate AICc for the GARCH (1,1) model, using q=2 in the formula for AICc. Comment on the statistical significance of the parameter values of your selected model. Write the complete form of the ARCH or GARCH model you have selected. Hand in the R output for the selected model, that is, the result of both summary(model) and logLik(model) but only for the one model that was selected by AICc. Also evaluate the unconditional (marginal) variance of the shocks in this model.**

For ARCH(0) log likelihood manual calculation, result is as follows,

```
> -0.5 * 1257 * (1 + log(2 * pi * mean(x^2)))
[1] 3065.963
```

Also, after I calculated the log likelihood of ARCH (q) by R, I calculated AICc by the below formula.

$$AIC_C = -2\log\text{likelihood} + 2(q+1)\frac{N}{N-q-2}$$

	q	logLik	N	AICc
GARCH(0,0)	0	3065.963	1257	-6129.9228
GARCH(0,1)	1	3125.785	1257	-6247.5604
GARCH(0,2)	2	3155	1257	-6303.9808
GARCH(0,3)	3	3182.452	1257	-6356.8721
GARCH(0,4)	4	3194.444	1257	-6378.84
GARCH(0,5)	5	3206.665	1257	-6401.2628
GARCH(0,6)	6	3203.627	1257	-6393.1643
GARCH(0,7)	7	3211.489	1257	-6406.8626
GARCH(0,8)	8	3217.258	1257	-6416.3717
GARCH(0,9)	9	3213.24	1257	-6406.3034
GARCH(0,10)	10	3218.855	1257	-6415.498

	q	logLik	N	AICc
GARCH(1,1)	2	3245.914	1257	-6485.8088

```
Call:
garch(x = x, order = c(1, 1), trace = F)

Model:
GARCH(1,1)

Residuals:
      Min        1Q    Median        3Q       Max
-4.216231 -0.590953 -0.006131  0.605057  4.972366

Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 6.151e-06   1.652e-06    3.724 0.000196 ***
a1 1.028e-01   1.353e-02    7.598 3.02e-14 ***
b1 8.867e-01   1.280e-02   69.295  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Diagnostic Tests:
        Jarque Bera Test

data:  Residuals
X-squared = 79.83, df = 2, p-value < 2.2e-16


        Box-Ljung test

data:  Squared.Residuals
X-squared = 1.4946, df = 1, p-value = 0.2215

> logLik(model)
'log Lik.' 3245.914 (df=3)
```

GARCH(1,1) result is as follows, and we get a log likelihood of 3245.914. By running the above formula for AICc and setting q=2, I got AICc = -6485.808861. I also found that GARCH(1,1) is preferred by AICc, as it resulted in the lowest AICc, and will use this model for my further analysis.

I can see that all parameter values show statistical significance, showing 0.000196 for a0, 3.02e-14 for a1, and less than 2e-16 for b1.

Complete form of the model is as follows:

$h_t = 6.151e\text{-}06 + 1.028e\text{-}01(\varepsilon_{t-1})^2 + 8.867e\text{-}01 h_{t-1}$, *where* h_t = *conditional variance*

Also, **unconditional variance** of the shocks in this model is followed by the formula :

w/(1-alpha-beta) = *6.151e-06 / (1 - 1.028e-01 - 8.867e-01) = 0.00058580952*

5) **Using the Minitab output from problem 2, and the R output from your selected model in problem 4, construct a 95% one step ahead forecast interval for the log exchange rate, based on your ARIMA-ARCH model. Compare this to the interval based on the ARIMA only model from problem 2. Also compute the 5th percentile of the conditional distribution of the next period's log exchange rate.**

.

Ans)

Using my ARIMA-ARCH model, I first used conditional variances from R. Using the result from problem 6, I get conditional variances h_t

Using h_{1258} = 0.000339766531529886 and ε_{1258} = *0.0084062467445901,*

I can get h_{1259} = 0.00031468634389677 based on the formula

$h_t = 6.151e\text{-}06 + 1.028e\text{-}01(\varepsilon_{t-1})^2 + 8.867e\text{-}01 h_{t-1}$, *where* h_t = *conditional variance*

Using h_{1259} = 0.00031468634389677, best one-step forecast $f_{1258,1}$ = 4.65156, and the formula,

$$f_{t,1} \pm 1.96\sqrt{h_{t+1}}.$$

I can get 95% intervals of (4.61679077425777, 4.68632922574223).

From problem 2, I have below forecast with 95% forecast interval.

Forecasts from Time Period 1258

			95% Limits		
Time Period	Forecast	SE Forecast	Lower	Upper	Actual
1259	4.65156	0.0211512	4.61010	4.69303	

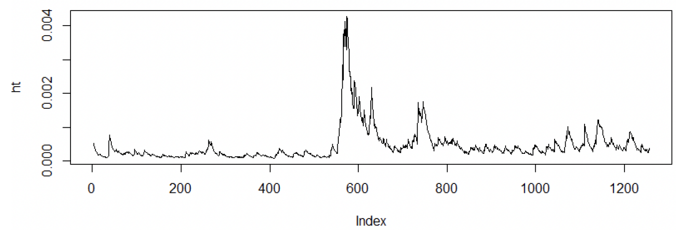
Comparing the above results with ARIMA forecast interval (4.61010, 4.69303), ARIMA-ARCH model shows more narrow intervals.

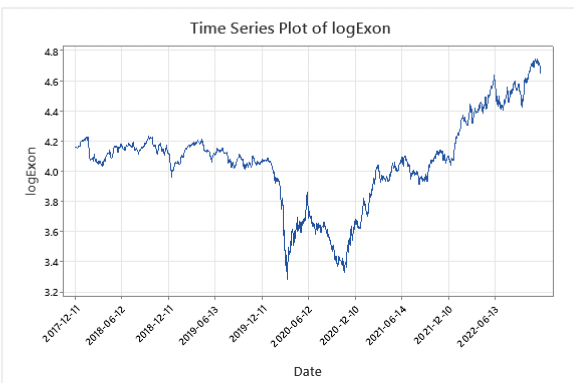
Also, the 5th percentile of the conditional distribution of the next period's log exchange rate is

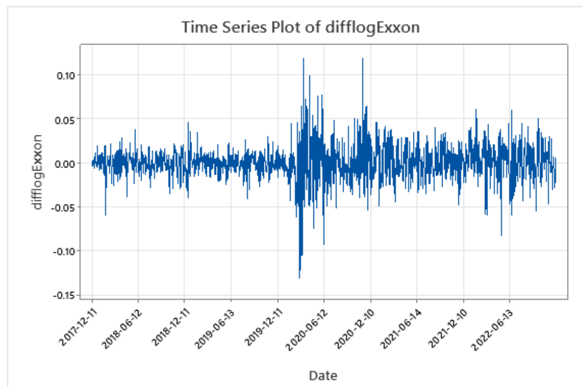
$f_{1258,1} - 1.645\sqrt{h_{1259}}$ = 4.62237868553778

6) **Plot the conditional variances, ht, for your fitted ARCH model from problem 4. Use this plot to locate bursts of high volatility. Do these highly volatile periods agree with those found from examination of the time series plot of the log exchange rates themselves?**

Ans)
By comparing highly volatile periods, which occur around the index of 574 or March 2020, the two highly volatile periods agree with those found from examination of the time series plot of the log stock prices.

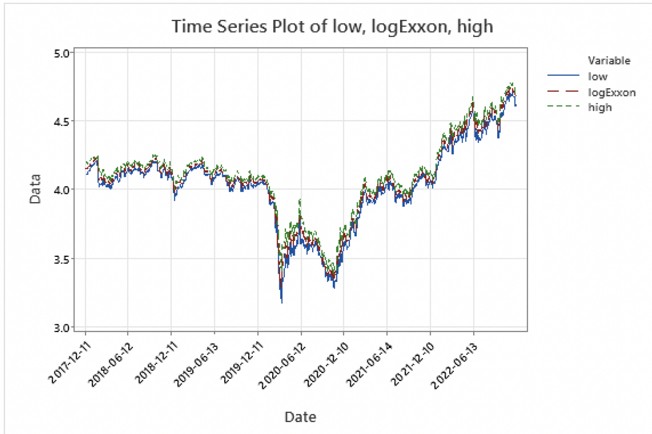






7) **Make a time series plot which simultaneously shows the log exchange rates, together with the ARIMA-ARCH one-step-ahead 95% forecast intervals based on information available the previous day. Using the plot, together with the numerical values in your Minitab worksheet, comment on the accuracy and practical usefulness of the forecast intervals. Keep in mind that the performance may be somewhat better here than in an actual forecasting context, since the ARIMA-ARCH parameters are estimated from the entire dataset, not just the observations up to the time at which the forecast is to be constructed.**

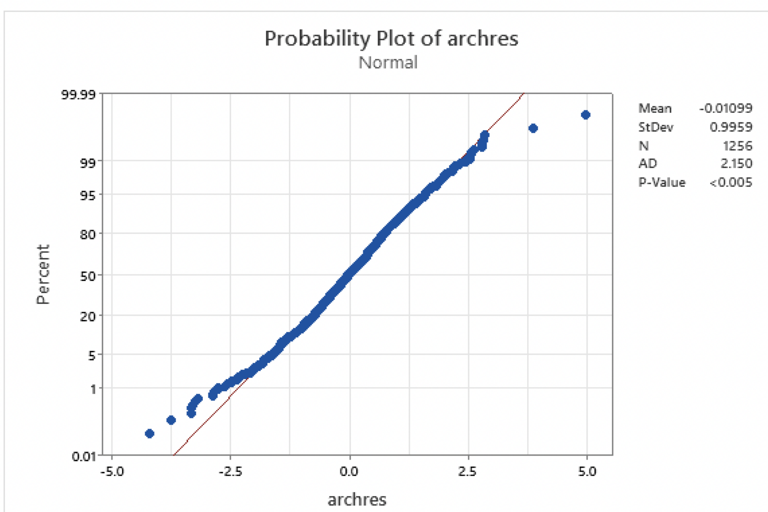
Ans)
ARIMA-ARCH model's accuracy seems better, showing a narrower interval compared to ARIMA only model. Forecast should have high accuracy, as it is just a one step ahead prediction. Also, as conditional volatility is relatively low, we may expect to have better accuracy. Meanwhile, we may expect to have lower accuracy, if we predict more steps ahead.
For practical usefulness of the forecast intervals, ARIMA-ARCH model is useful as the model's parameters are estimated from the entire data set with the conditional variances that keep updated over the entire course of period. This is of more practical interest to the forecaster than the volatility of the series considered as a whole.



8) **Compute the residuals from your ARIMA-ARCH model, that is e_t = E_t / h_t^(0.5). If the ARIMA-ARCH model is adequate, these residuals should be normally distributed with mean zero and variance 1. Make a normal probability plot of archres. Does the model seem to have adequately described the leptokurtosis ("long-tailedness") in the data?**

Ans)

The model seems to have long-tail, which seems to fail the normality test. Therefore, it adequately describes the leptokurtosis in the data. I've also checked the p-value, and the value is <0.005. This also gives evidence to reject the null hypothesis of normality test, indicating a non-normal distribution.



9) **From the formula for the prediction intervals, it follows that the 95% prediction interval constructed yesterday fails to cover today's log exchange rate whenever today's residual exceeds 1.96 in absolute value. What percentage of the time did the intervals fail?**

Ans)

Based on the calculation, there are 69 failures, and therefore for a total 1,256 data values in archres, the percentage is 5.49363057325 %.

Final question

Check whether either or both of the one-step-ahead forecast intervals calculated in part 5 actually contained the n+1 st observation. Based on this, does the ARMA only interval seem too wide, too narrow, or just about right? Then answer the same question for the ARMA-ARCH interval.

Ans)

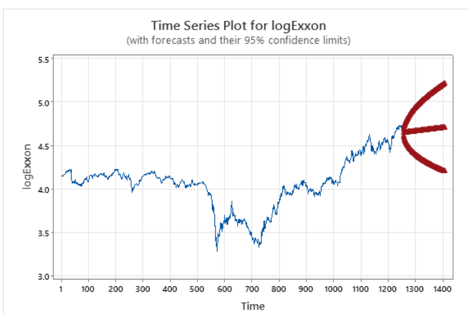
Both of the one-step-ahead forecast intervals calculated in part 5 actually contained the n+1 st observation.

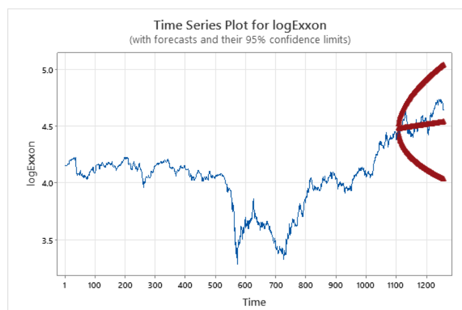
- one-step-ahead forecast intervals calculated in part 5: (4.61679077425777, 4.68632922574223)
- n+1 st observation: 4.639958021131858

For ARMA only intervals, I've conducted back testing. As you can see, intervals from back testing includes all the future observations, and therefore shows just about the right range. (There are a couple of adjacent points for upper interval and point observations) In terms of the lower bound, it may seem a bit wider, but overall, the width of the interval seems reasonable.

Also, comparing unconditional variance with conditional variance, unconditional variance is higher than conditional variance, stating that ARMA only intervals might be wider compared to ARMA-ARCH intervals.

- ARMA only forecast interval (4.61010, 4.69303)
- Unconditional variance = 0.00058580952, h_1259 = 0.00031468634389677





For ARMA-ARCH intervals, we need to compare the interval with conditional volatility h_t. As you can see, the conditional volatility in n+1 st point is relatively lower compared to the peak point in the year 2020, and also lower than the mean of h_t. Therefore, we can say that ARMA-ARCH intervals are a bit narrow.

- ARMA-ARCH forecast interval (4.61679077425777, 4.68632922574223)
- Mean of h_t = 0.0004602062515897345, h_1259 = 0.00031468634389677