Riazul Islam

May 6, 2020

Forecasting Time Series Data

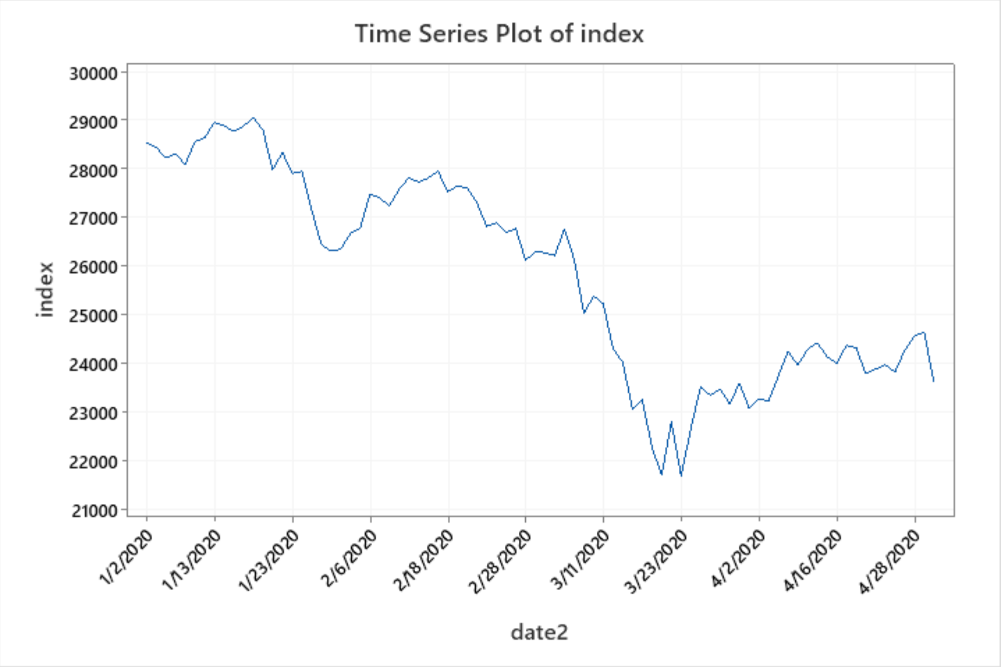
Project 2

The Hang Seng Index (HSI) is a measurement of the daily changes of the largest companies of the Hong Kong Stock Exchange and is the main indicator of the overall market performance in Hong Kong.[1] 50 companies make up the Hang Seng, representing 58% of the capitalization of the Hong Kong Stock Exchange and many of the leading companies in Hong Kong and China.[2] As a result, it can be a strong indicator of the perceived financial strength of both the Hong Kong Special Administrative Region and of the People's Republic of China.

Due to the coronavirus outbreak, the Hang Seng has experienced significant and varying volatility in the past 4 months. It is this data that will be analyzed in this project.

1) Plot the logs of <u>Hang Seng index</u>. Based on this 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?

Time series of Hang Seng index

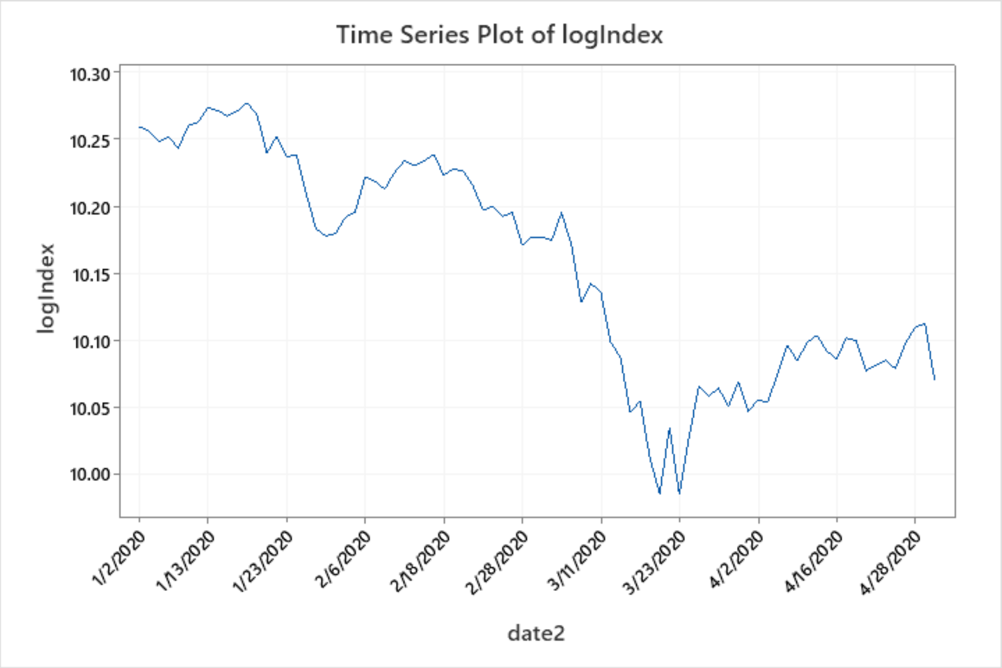


Since the start of 2020, the Hang Seng has declined quite a bit due to the coronavirus outbreak. Though the Hang Seng is measured in points rather than Hong Kong dollars, with the base of 100 points "set equivalent to the stocks' total value as of the market close on July 31, 1964", it would be wise to use the logged Hang Seng Index instead of just the index points.

[1] https://en.wikipedia.org/wiki/Hang_Seng_Index

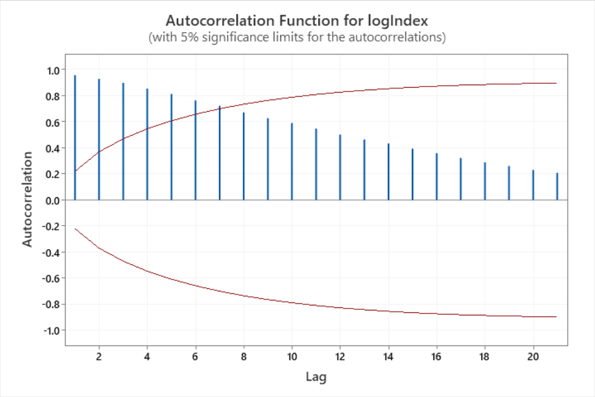
[2] https://www.hsi.com.hk/eng/indexes/all-indexes/hsi, https://en.wikipedia.org/wiki/Hang_Seng_Index

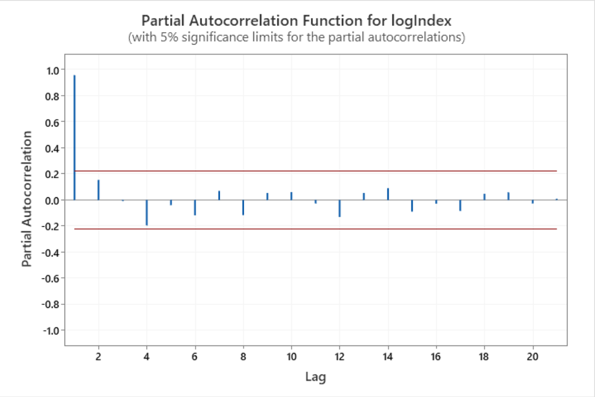
Time series of logged Hang Seng index



The logged Hang Seng index plot looks basically the same as the unlogged index, but this will still be used for the analysis.

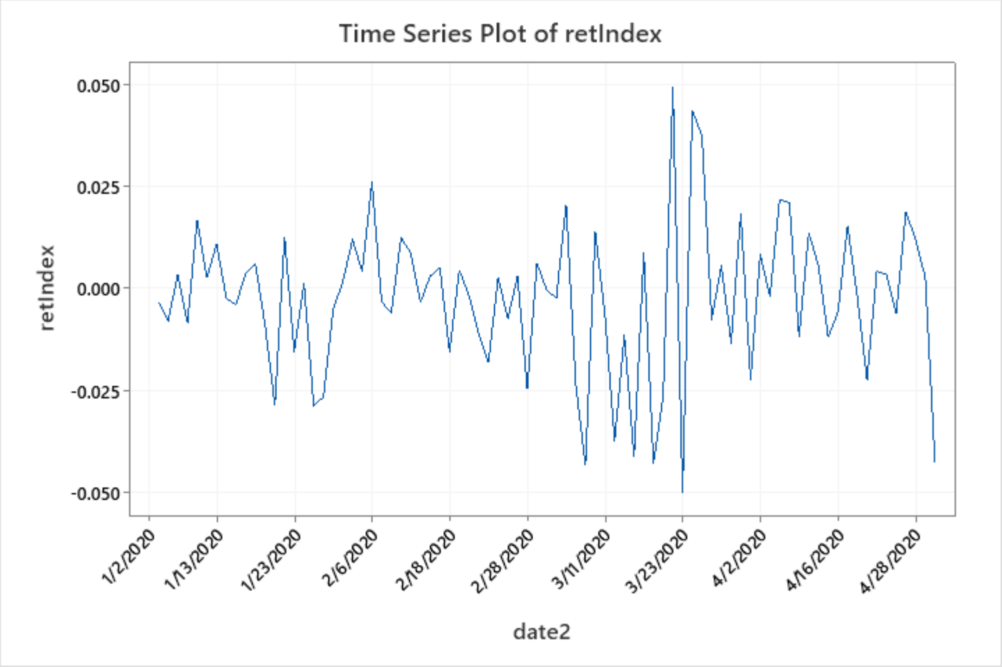
ACF & PACF of logged Hang Seng index



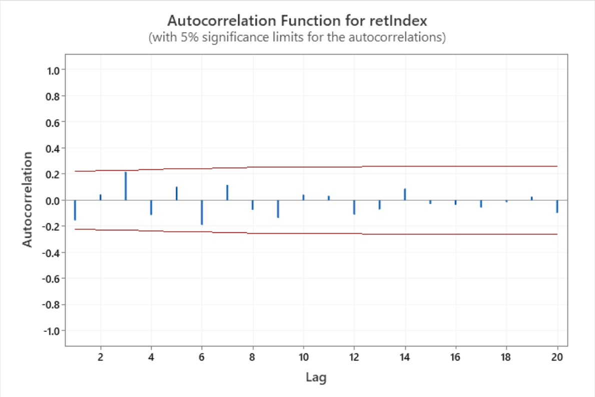


With the logIndex neither mean-reverting nor stationary, we should look at the differenced logged Hang Seng index, equivalent to the returns. If this version of the data was intended for use, since the ACF dies down and the PACF cuts off it looks like an appropriate ARMA model could be (1,0). However, the returns on the Hang Seng, or the differenced log Hang Seng, should also be considered before moving forward.

<u>Time series of returns on Hang Seng index</u>

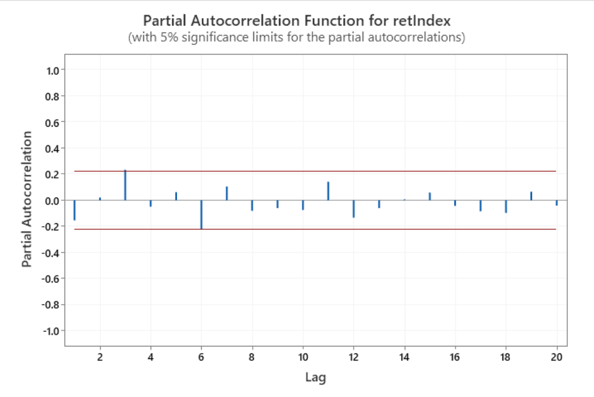


<u>ACF & PACF of returns on Hang Seng index</u>



Autocorrelation Function for retIndex

(with 5% significance limits for the autocorrelations)



Partial Autocorrelation Function for retIndex

(with 5% significance limits for the partial autocorrelations)

The returns on the Hang Seng do appear to be stationary and mean-reverting, so it will be used for this analysis. Based on the ACF and PACF of retIndex, an appropriate ARIMA model may be (0,1,0), pointing toward a random walk. However, this part of the analysis is useful for identifying the d metric to use, and the AIC_C from ARIMA models without constants will be used to determine the p and q to use for modeling the AR and MA components.

2) Using AIC$_C$, select an ARIMA (p, 1, q) (without constant) with $0 \leq p \leq 2$, $0 \leq q \leq 2$. 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 the (ARIMA) one step ahead forecast and 95% forecast interval.

P	D	Q	SS	AIC	AICc
0	1	0	0.028916	-642.962	-640.911
1	1	0	0.028312	-644.673	-640.519
2	1	0	0.028274	-644.782	-638.47
0	1	1	0.028395	-644.437	-640.283
0	1	2	0.027823	-646.085	-639.773
1	1	1	0.028304	-644.697	-638.385
1	1	2	0.026533	-649.93	-641.403
2	1	1	0.027777	-646.219	-637.692
2	1	2	0.026721	-649.357	-638.557

Based on the AIC_C, the ARIMA(1,1,2) model without constant should be selected since it has the lowest AIC_C.

Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	-0.764	0.115	-6.66	0.000
MA 1	-0.754	0.136	-5.56	0.000
MA 2	0.187	0.123	1.53	0.130

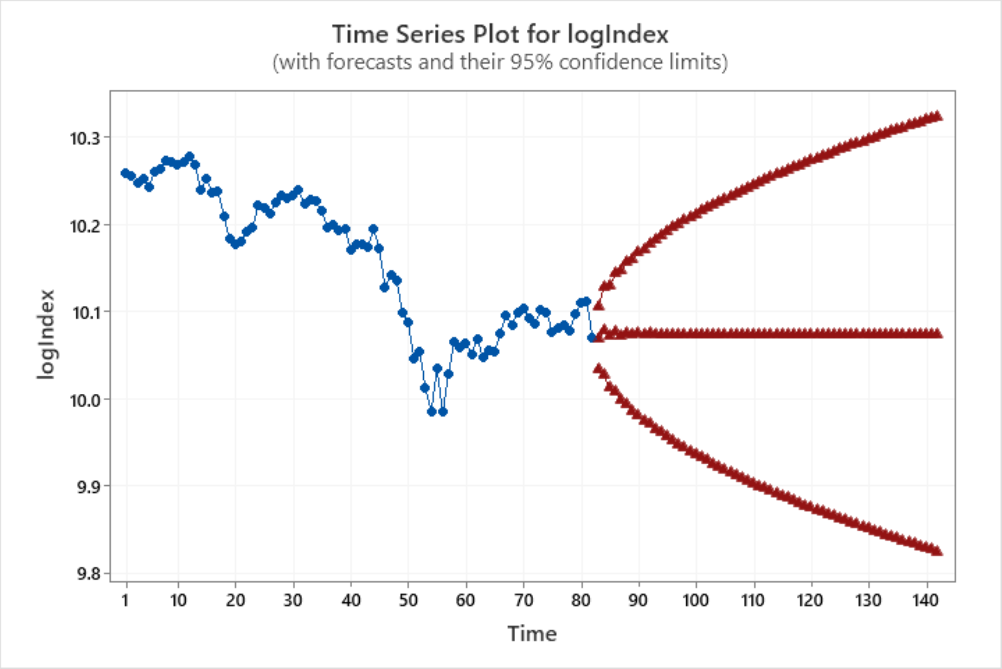
The complete form of the fitted model is:

$$x_t = -0.764x_{t-1} + 0.754\varepsilon_{t-1} - 0.187\varepsilon_{t-2} + \varepsilon_t$$

The one-step ahead forecast will be:

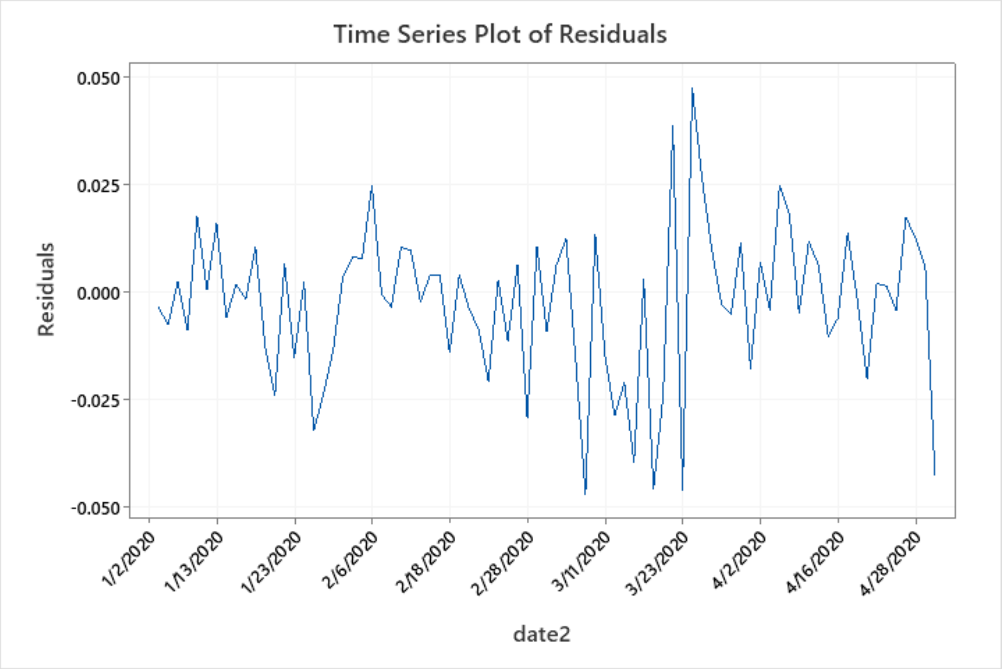
Forecasts from period 82

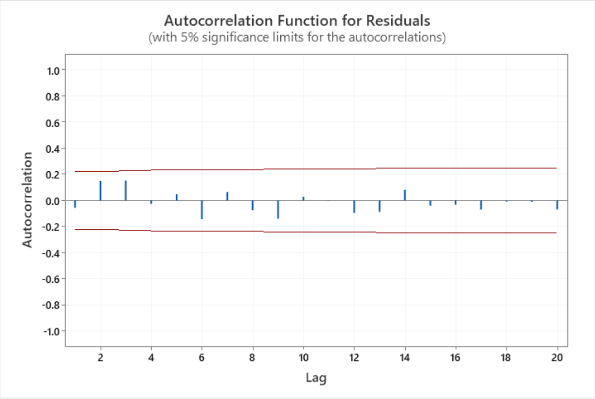
		95% Limits		
Period	Forecast	Lower	Upper	Actual
83	10.0689	10.0327	10.1051	

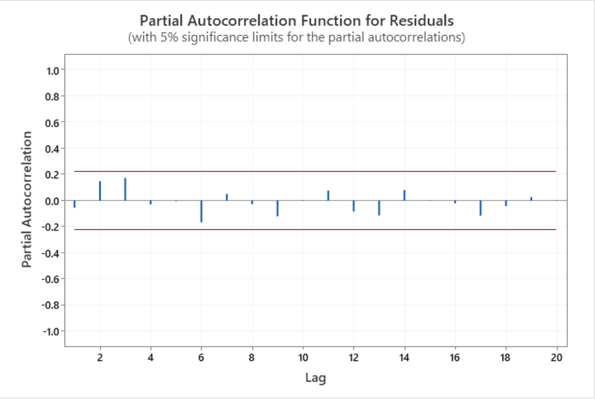


3) Plot the residuals, as well as ACF and PACF of both the residuals and the squared residuals. Use these plots to argue that the residuals, although approximately uncorrelated, are not independent; instead, they show evidence of conditional heteroscedasticity.

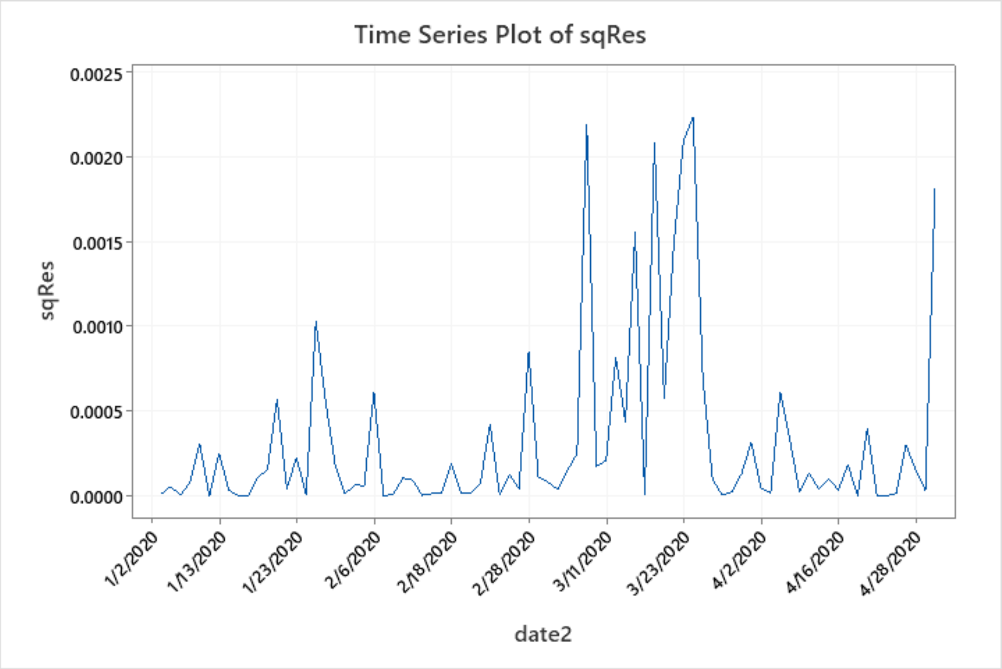
Residual Plots

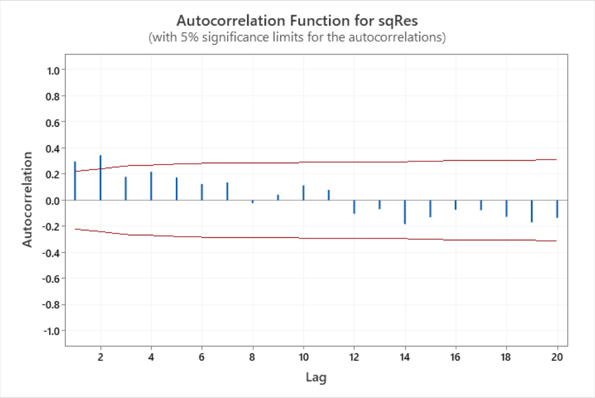


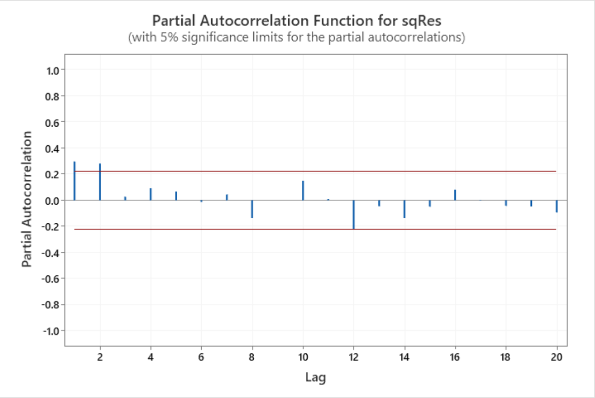




Squared Residuals



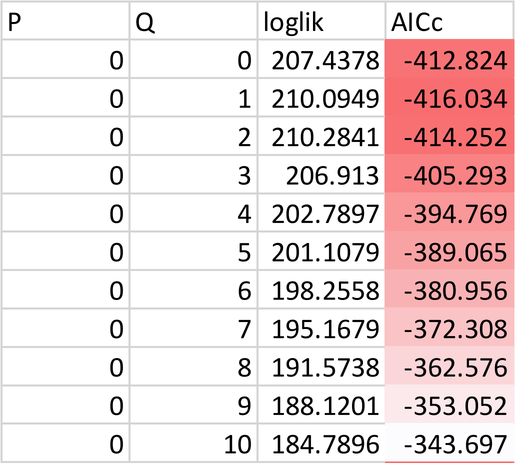




These plots show that there is high clustering of large squared errors. This demonstrates conditional heteroscedasticity: the magnitude of errors suggests the magnitude of future errors. Since residuals seems to bounce a lot and seem to be uncorrelated, while squared residuals tend to cluster, there is evidence of conditional heteroscedasticity.

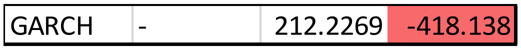
4) Using R on the residuals from the ARIMA model, find the log likelihood values and AICC values for ARCH(q) models where q ranges from 0 to 10. You will need to calculate the log likelihood for the ARCH(0) model by hand. See the handout on Estimation and Automatic Selection of ARCH models.

ARCH(q) models



P	Q	loglik	AICc
0	0	207.4378	-412.824
0	1	210.0949	-416.034
0	2	210.2841	-414.252
0	3	206.913	-405.293
0	4	202.7897	-394.769
0	5	201.1079	-389.065
0	6	198.2558	-380.956
0	7	195.1679	-372.308
0	8	191.5738	-362.576
0	9	188.1201	-353.052
0	10	184.7896	-343.697

GARCH(1,1) model



GARCH	-	212.2269	-418.138

Since the GARCH(1,1) model has the lowest AIC_C among all tested models, this will be chosen as the selected model.

```
Model:
GARCH(1,1)

Residuals:
       Min         1Q     Median         3Q        Max 
-2.8933909 -0.6966373 -0.0001442  0.4813443  1.6541344 

Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 6.316e-05   7.225e-05    0.874    0.382
a1 2.183e-01   2.337e-01    0.934    0.350
b1 6.037e-01   3.813e-01    1.583    0.113

Diagnostic Tests:
        Jarque Bera Test

data:  Residuals
X-squared = 5.004, df = 2, p-value = 0.08192


        Box-Ljung test

data:  Squared.Residuals
X-squared = 0.39136, df = 1, p-value = 0.5316
```

The ω, α, and β *p*-values are all not significant at the 95% level. The Jarque Bera Test *p*-value is just above 5%, which means we may consider accepting the null hypothesis that the residuals are normally distributed and that the model may not demonstrate leptokurtosis. The Ljung-Box *p*-value is large, and so we do not reject the null hypothesis (no autocorrelation in the squared residuals).

The complete form of the selected GARCH(1,1) model is:

$$\varepsilon_t \mid \psi_{t-1} \sim N(0, h_t)$$

$$h_t = (6.316 \times 10^{-5}) + 0.2183\varepsilon_{t-1}^2 + 0.6037h_{t-1} + \varepsilon_t$$

The unconditional variance of the shocks in this model is

$$\frac{\omega}{(1-\alpha-\beta)} = 0.0003548315 = \ 3.548315 \times 10^{-4}$$

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. (If you decided to use a GARCH(1,1) model, you will need to first get the conditional variances from R. See Problem 6.) 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.

The 95% one-step ahead forecast interval for the log exchange rate based on the ARIMA(1,1,2)-GARCH(1,1) model would be:

$$h_t = (6.316 \times 10^{-5}) + 0.2183\varepsilon_{t-1}^2 + 0.6037h_{t-1} + \varepsilon_t$$

$$h_{t+1} = (1.531 \times 10^{-7}) + 0.2183(-0.04268)^2 + 0.6037(2.18 \times 10^{-4}) + (0) = 5.294 \times 10^{-4}$$

$$f_{t,1} \pm 1.96\sqrt{h_{t+1}} = f_{t,1} \pm 1.96\sqrt{5.294 \times 10^{-4}} = f_{t,1} \pm 0.0230089$$

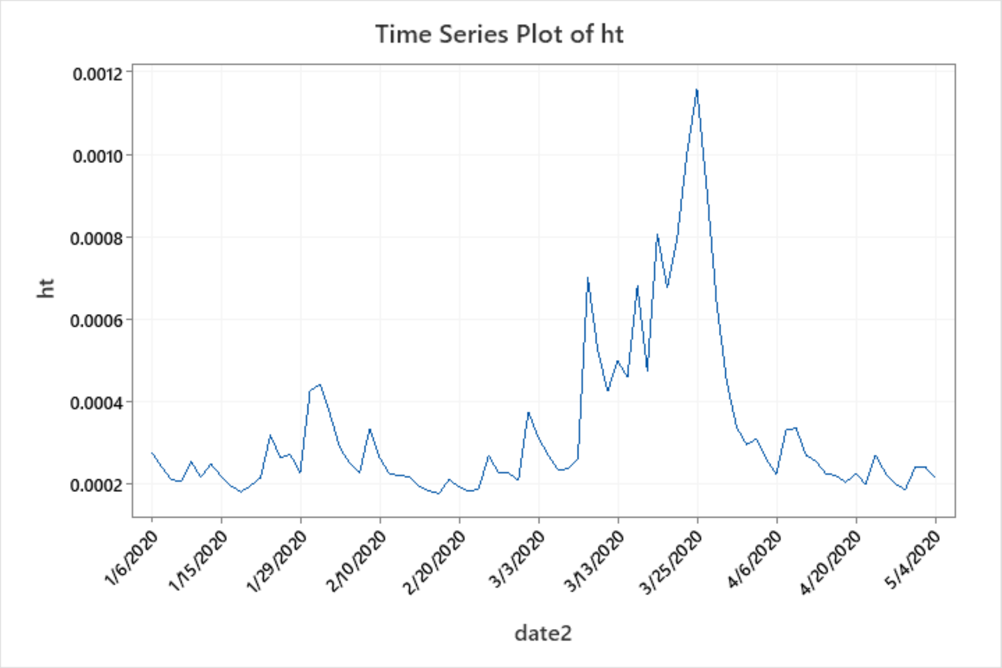
$$forecast\ interval = 10.02119 \pm 0.0230089$$

The 95% confidence interval is between (10.02119, 10.11661).

The 95% interval is slightly wider than the ARIMA only model's interval from question 2. This is because variance in the latest period (t=81 to t=82) is quite high compared to the whole dataset, so the conditional variance drives the one-step ahead forecast confidence interval to be slightly wider.

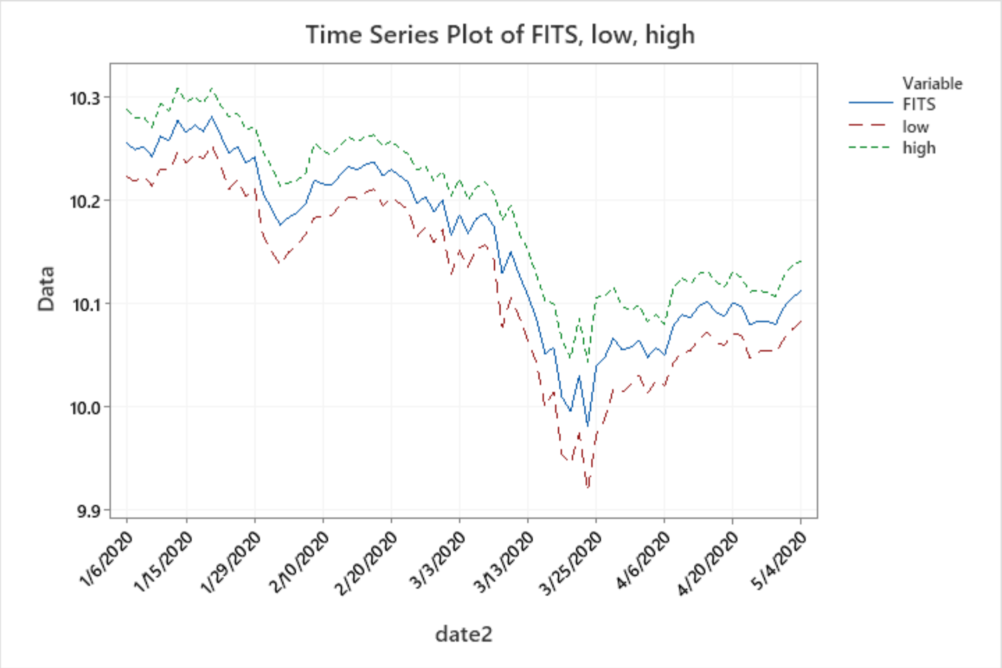
The 5th percentile of the conditional distribution of t=83's logIndex is 10.02886, which is slightly larger than the lower bound of the 95% ARIMA(1,1,2)-GARCH(1,1) forecast interval.

6) Plot the conditional variances, h_t, for your fitted ARCH model from problem 4. (See instructions below). 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?



The period of highest volatility in this graph, during the month March, matches the period of highest volatility identified in the time series plot of the logged Hang Seng Index.

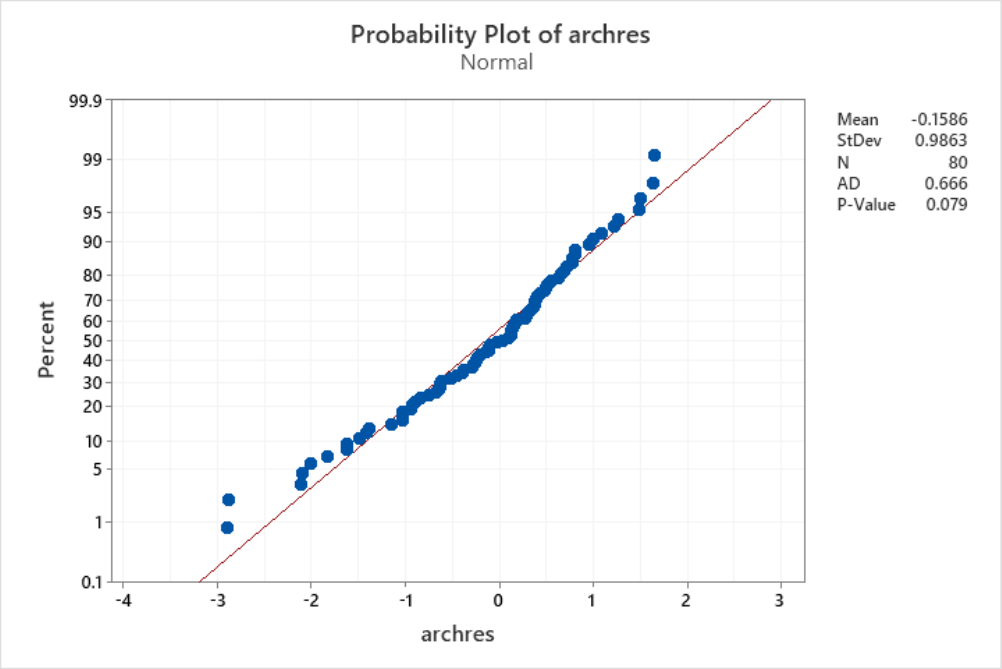
7) Make a time series plot which simultaneously shows the log exchange rates, together with the ARIMA-ARCH one-step-ahead 95% forecast intervals based on information available the previous day. (See instructions below). Using the plot, together with the numerical values in your Minitab worksheet, comment 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 data set, not just the observations up to the time at which the forecast is to be constructed.



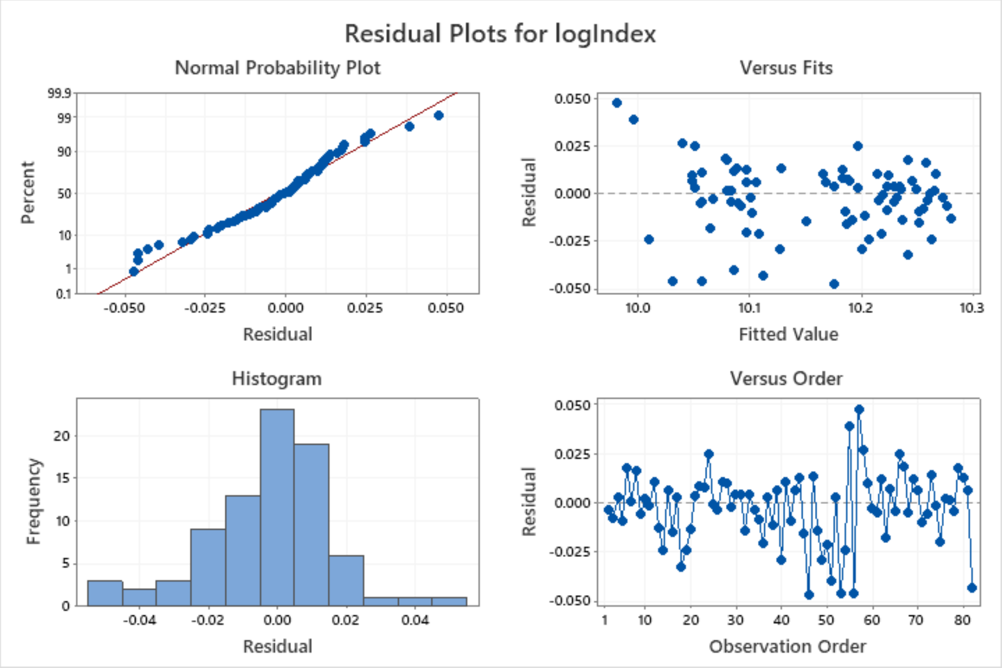
The logged Hang Seng index falls within the forecast interval throughout the dataset, though there are periods of severe drops when it almost falls outside the one-step ahead forecast. It does not look that way when rising, which may be a sign that the most volatile moves are downwards, followed by a rise that already takes into account the previous volatility from the downward movement. These forecast intervals also widen in times of high volatility. They are somewhat useful, but given that the same data is being used to both calculate forecast intervals and test those forecast intervals, they are not of the most practical use.

8) Compute the residuals from your ARIMA-ARCH model, that is, $e_t = \varepsilon_t / \sqrt{h_t}$. If the ARIMA-ARCH model is adequate, these residuals should be normally distributed with mean zero and variance 1. To compute these residuals in Minitab, use Calc → Calculator → Store result in variable: archres, Expression: RESI1/sqrt(ht). Make a normal probability plot of archres, using Stat → Basic Statistics → Normality Test. Does the model seem to have adequately described the leptokurtosis ("long-tailedness") in the data?

Normal Probability Plot of archres using ARIMA(1,1,2)-GARCH(1,1) model



Residual Plots for logIndex using ARIMA(1,1,2) model



When comparing the normality plot of the ARIMA(1,1,2)-GARCH(1,1) model with the ARIMA(1,1,2) model, the former seems to better account for leptokurtosis in the data. However, the mean is not very close to 0 in the archres normal probability plot.

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. Use Calculator to count up how many failures there were, using sum(abs(archres)>1.96). What percentage of the time did the intervals fail? (Keep in mind that there are not 1259 data values in archres).

There was a failure rate of 6.25% (5 failures out of 80).

Finally, 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.

The Hang Seng index on May 5th, 2020 (t=83) was 23,868.7, which means that the logged Hang Seng Index was 10.08032. This is well within the ARIMA-only 1-step ahead 95% confidence interval (10.0327, 10.1051) and the ARIMA-GARCH 1-step ahead 95% confidence interval (10.02119, 10.11661). This initially would make it seem like both intervals are quite wide, but I believe they are just about right given the recent increase in volatility due to the steep decline in the Hang Seng the day before, dropping from 24,643.6 to 23,613.8 from t=81 to t=82.