Forecasting Time Series Data (STAT-GB - 6018)
Professor Clifford Hurvich
PROJECT 2:

Predicting Bitcoin Stock Prices using ARIMA-ARCH Models

1) Outline

My chosen dataset is sourced from Kaggle, and contains **2836 daily observations** of **Bitcoin prices in USD**. For this project, I will focus on the *Date* and *Adjusted Close* variables — where the Date represents the daily timestamp, and Adjusted Close reflects the adjusted closing price, accounting for market factors.

```
[2836 rows x 7 columns]
             Adj Close
Date
2017-01-01    998.325012
2017-01-02   1021.750000
2017-01-03   1043.839966
2017-01-04   1154.729980
2017-01-05   1013.380005
...                  ...
2024-10-02  60632.785156
2024-10-03  60759.402344
2024-10-04  62067.476562
2024-10-05  62089.949219
2024-10-06  62818.953125
[2836 rows x 1 columns]
```

(↑ Original Dataset)

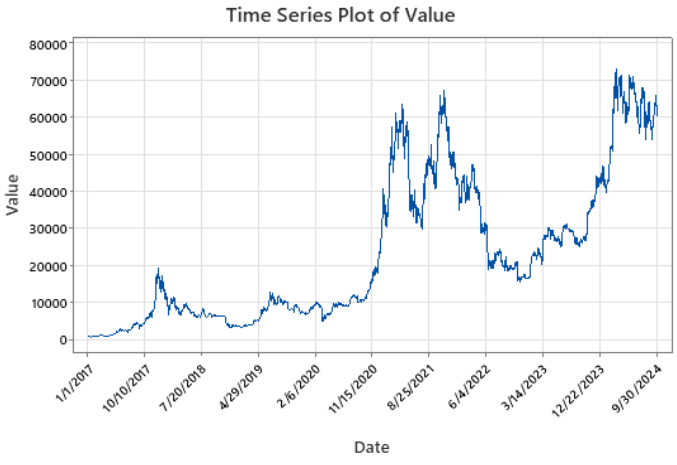
	C1-D	C2
	Date	Value
2833	10/3/2024	60759.4
2834	10/4/2024	62067.5
2835	10/5/2024	62089.9
2836		

(↑ After removing most recent observation)

The most recent observation (October 6, 2024) will be excluded and used to evaluate the forecast model's performance.

2) Preparing the data for analysis

a) Check for stationarity:

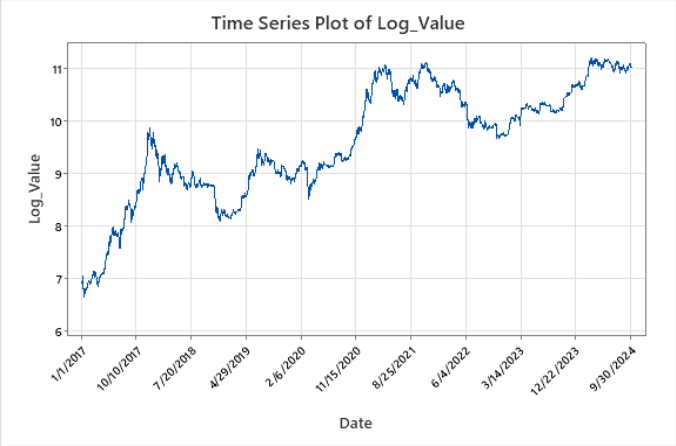


The original time series plot of Bitcoin's value implies non-stationary behaviour. Both the mean and the variance appear to change with time.

The plot also exhibits extreme fluctuations, with sharp upward and downward movements reflecting the cryptocurrency's exponential growth and high volatility, which makes it challenging to model directly due to large variations in the scale of the values.

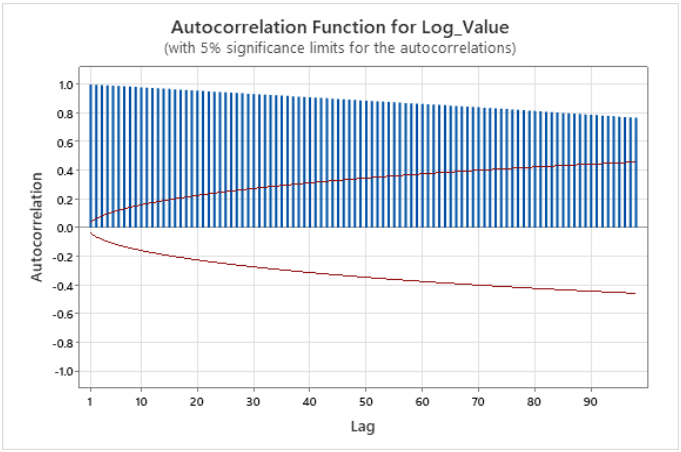
Although not shown for the sake of preventing cluttering this page with unessential information — the raw data's ACF plot revealed a gradual decay, while its PACF had significant spikes — confirming the non-stationarity of the series and the need for a transformation that would make the series more suitable for analysis.

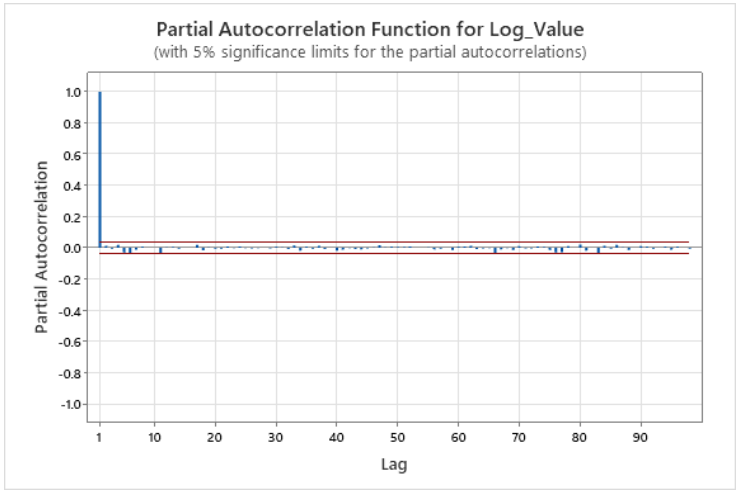
b) Log Transformation:



The log-transformed series successfully reduced the variability in the data, however it still seemed to remain non-stationary.

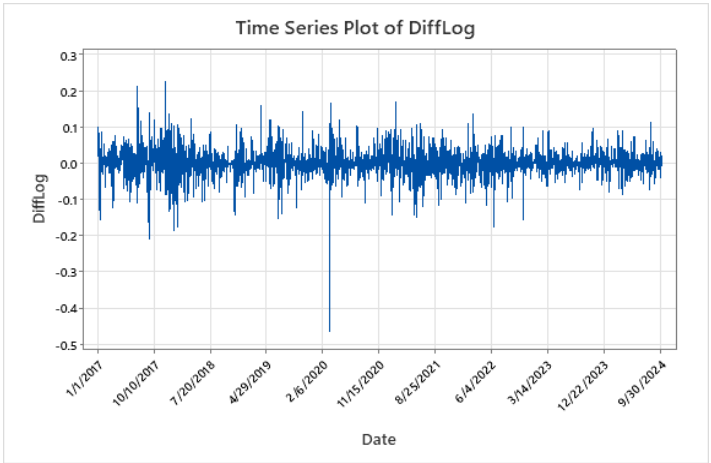
This was evident in the ACF plot, which displayed a slow decay rather than a sharp cutoff, and the PACF, which at lag 1 is very close to 1 — which is a sign of a random walk, and thus a sign that the series is indeed non-stationary, implying the need for differencing.

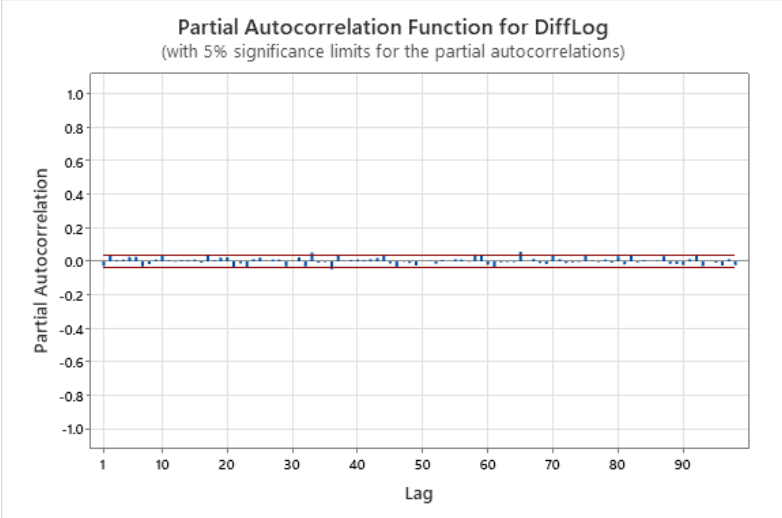


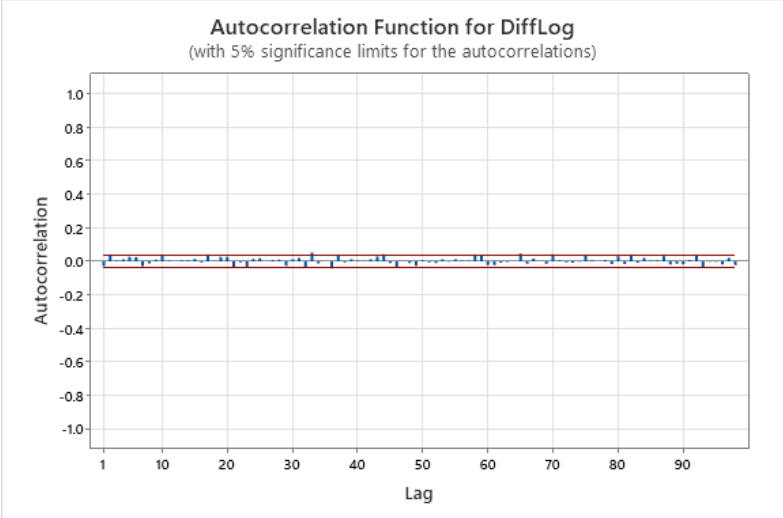


c) Applying differencing:

After applying the **first-order differencing** to the log-transformed series, the data now appears stationary, as seen in the Time Series Plot of it below.







The ACF quickly tapers off, which is typical for a stationary series. The values also mostly fall within the confidence interval bands from the initial lags, suggesting that the series is now stationary. The PACF has subsequent lags mostly within the confidence interval. The significant spikes at lag 1 in both ACF and PACF suggest that the data already appears to be stationary after one differencing. Further differencing is not necessary, as it could risk being prone to over-differencing and overfitting to short-term variations instead of preserving the overall dynamics of the time series.

Therefore, an ARIMA model with **d=1** (i.e. first-order differencing) seems appropriate.

d) Preliminary inferences on potential parameters for an **ARIMA(p,d,q)** model:

As for the remaining 'p' and 'q' parameters of the ARIMA model, we can make inferences from looking at the above plots of the first-order differenced, log-transformed series.

The PACF shows significant spikes at **lags 1 and 2**, suggesting the presence of **two autoregressive components**, and indicating that the autoregressive component **'p' might be 2**. The ACF does not show significant spikes outside the confidence bounds beyond the initial lags, suggesting the **absence of moving average components**, and indicating that the moving average component **'q' might be 0**.

Hence, based on the plots, we can conclude that an **ARIMA(2,1,0)** model *may* be appropriate, although further validation will be needed to confirm its optimality.

3) Fitting the data to an ARIMA model

a) Model Selection:

According to the following Minitab output, the ARIMA(2,1,0) model with a constant term included is identified as the optimal model based on the lowest AICc value.

ARIMA models with constant term:

Model Selection

Model (d = 1)	LogLikelihood	AICc	AIC	BIC
p = 2, q = 0*	5250.1	-10492.3	-10492.3	-10468.5
p = 0, q = 2	5250.13	-10492.3	-10492.3	-10468.5
p = 2, q = 1	5250.82	-10491.6	-10491.6	-10461.9
p = 1, q = 2	5250.70	-10491.4	-10491.4	-10461.6
p = 2, q = 2	5251.61	-10491.2	-10491.2	-10455.5
p = 1, q = 1	5249.39	-10490.8	-10490.8	-10467.0
p = 1, q = 0	5248.19	-10490.4	-10490.4	-10472.5
p = 0, q = 1	5248.11	-10490.2	-10490.2	-10472.4
p = 0, q = 0	5247.02	-10490.0	-10490.0	-10478.1

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

ARIMA models without constant term:

Model Selection

Model (d = 1)	LogLikelihood	AICc	AIC	BIC
p = 2, q = 2*	5250.6	-10491.1	-10491.1	-10461.3
p = 2, q = 1	5249.48	-10491.0	-10491.0	-10467.2
p = 1, q = 2	5249.35	-10490.7	-10490.7	-10466.9
p = 2, q = 0	5248.11	-10490.2	-10490.2	-10472.4
p = 0, q = 2	5248.09	-10490.2	-10490.2	-10472.3
p = 1, q = 1	5247.22	-10488.4	-10488.4	-10470.6
p = 1, q = 0	5245.99	-10488.0	-10488.0	-10476.1
p = 0, q = 0	5244.94	-10487.9	-10487.9	-10481.9
p = 0, q = 1	5245.91	-10487.8	-10487.8	-10475.9

** 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	-0.0277	0.0188	-1.47	0.141
AR 2	0.0373	0.0188	1.98	0.047
Constant	0.001444	0.000713	2.02	0.043

Differencing: 1 Regular
Number of observations after differencing: 2834

b) Complete form of the fitted model:

$$x_t = 0.001444 - 0.0277x_{t-1} + 0.0373x_{t-2} + \epsilon_t$$

Where:

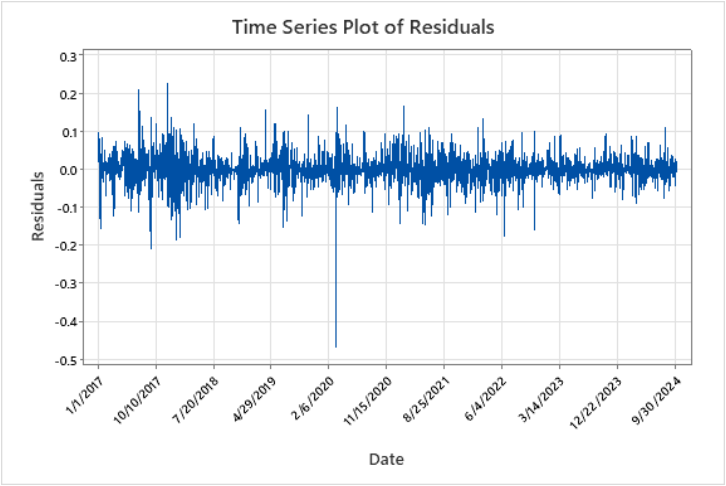
- x_t: The value of the time series at time t, representing the first-order differenced and log-transformed Bitcoin data.
- ϵ_t: The error term at time t, assumed to follow a white noise process with a mean of zero and constant variance.
- t: The time index, indicating the specific time period for each observation in the time series.

c) Minitab's "One Step Ahead" forecast of the ARIMA model with 95% forecast intervals:

Forecasts from Time Period 2835

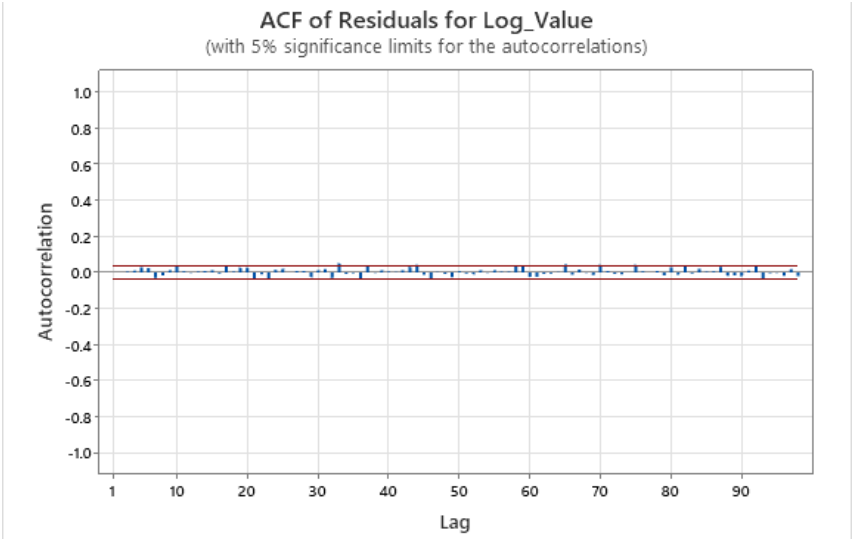
			95% Limits	
Time Period	Forecast	SE Forecast	Lower	Upper	Actual
2836	11.0386	0.0379696	10.9641	11.1130	

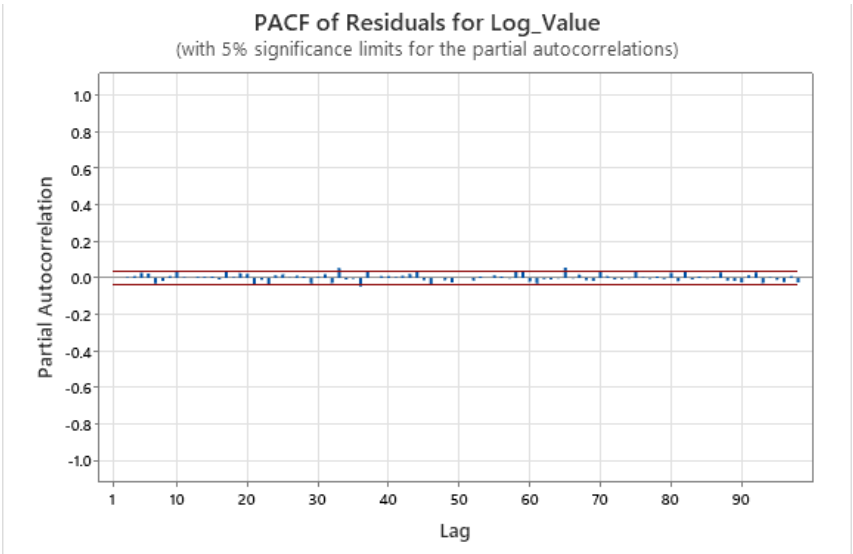
d) <u>The residuals:</u>



The residuals appear to be randomly distributed around zero with no discernible trend, suggesting that the ARIMA(2,1,0) model has successfully captured the serial correlations in the data.

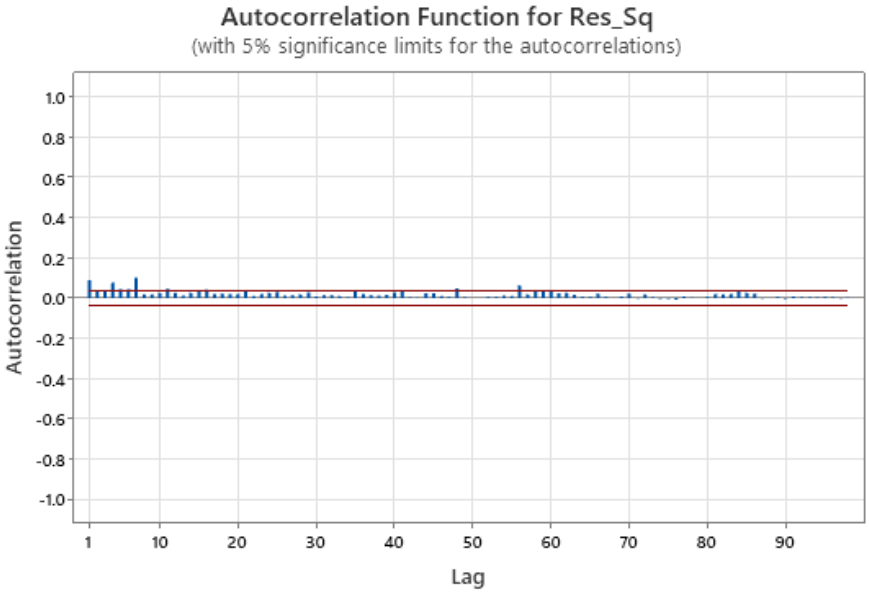
The ACF and PACF plots of the residuals below show no significant spikes beyond the confidence intervals, reinforcing the indication of no significant autocorrelations. Thus, we can conclude that the residuals themselves are approximately uncorrelated.

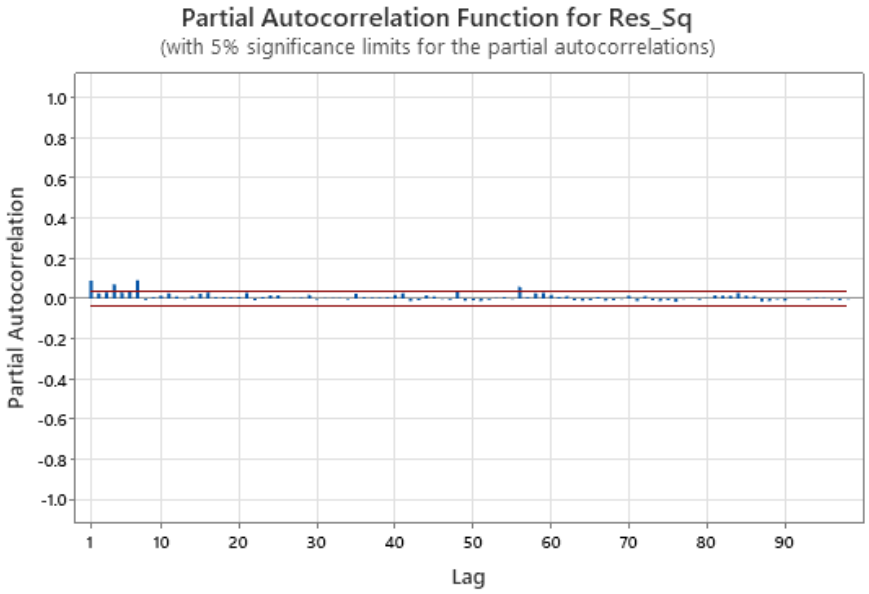




However, the *squared* residuals exhibit significant spikes in the ACF and PACF plots below, at early lags, suggesting autocorrelation in the variance of the residuals. This is a hallmark of "conditional heteroscedasticity", where the variance is not constant but changes over time.

The presence of autocorrelation in the squared residuals implies that the variance is dependent on past values, meaning the residuals are not independent. And looking at the ACF and PACF plots — indeed, when a shock or change in variance occurs, it tends to persist for a certain period following its initial occurrence.





This persistence of shocks or changes in variance over time is a common feature of conditional heteroscedasticity and is often better captured by models like ARCH or GARCH.

4) Selecting a model better suited for conditional heteroscedasticity, to fit the data to

a) Evaluating Autoregressive Conditional Heteroscedasticity (ARCH) models:

```
##      q   loglik      aicc
## 1    0 5250.158 -10500.32
## 2    1 5294.300 -10588.60
## 3    2 5309.529 -10619.05
## 4    3 5330.579 -10661.14
## 5    4 5421.077 -10842.13
## 6    5 5451.387 -10902.74
## 7    6 5475.030 -10950.02
## 8    7 5481.848 -10963.64
## 9    8 5484.802 -10969.54
## 10   9 5483.056 -10966.03
## 11  10 5484.183 -10968.27
```

The R Markdown output on the left shows the AICc values for ARCH(q) models ranging from q=0 to q=10.

The model with the lowest AICc value is ARCH(8), which has an AICc value of **-10,969.54**, making it the most suitable ARCH model based on the AICc criterion.

b) Evaluating General ARCH (GARCH) models:

```
## The AICc for the GARCH(1,1) model is: -10898.89
```

The GARCH(1,1) model was evaluated, and its AICc value was calculated as **-10,898.89** using q=2 in the AICc formula.

Since the AICc for the ARCH(8) model is lower than that of the GARCH(1,1) model, the **ARCH(8)** model is preferred.

5) Fitting the data to the chosen model: ARCH(8)

a) Summary:

```
## Call:
## garch(x = residuals, order = c(0, 8))
##
## Model:
## GARCH(0,8)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -9.002268 -0.444798 -0.004378  0.467599  6.439802
##
## Coefficient(s):
##     Estimate  Std. Error  t value Pr(>|t|)
## a0 5.035e-04   1.852e-05   27.186  < 2e-16 ***
## a1 1.241e-01   1.437e-02    8.637  < 2e-16 ***
## a2 6.009e-02   1.079e-02    5.569 2.57e-08 ***
## a3 1.570e-02   9.997e-03    1.571  0.11625
## a4 2.195e-01   7.522e-03   29.175  < 2e-16 ***
## a5 8.245e-02   1.104e-02    7.469 8.10e-14 ***
## a6 1.100e-01   1.591e-02    6.912 4.77e-12 ***
## a7 4.692e-02   1.451e-02    3.234  0.00122 **
## a8 5.861e-02   1.310e-02    4.474 7.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Diagnostic Tests:
##  Jarque Bera Test
##
## data:  Residuals
## X-squared = 3916.3, df = 2, p-value < 2.2e-16
##
##
##  Box-Ljung test
##
## data:  Squared.Residuals
## X-squared = 0.032631, df = 1, p-value = 0.8566
```

b) log-Likelihood:

```
## Log-Likelihood for ARCH(8): 5484.802
```

The log-likelihood value (5484.802) indicates a good model fit.

c) Parameter Estimates:

From the ARCH(8) output, we observe that most coefficients ($\alpha_1, \alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8$) have p-values of less than 0.001, indicating they are statistically significant at conventional levels. However, α_3 has a p-value of 0.116, suggesting that it is not significant.

d) Complete form of the ARCH(8) model:

$$\sigma_t{}^2 = \omega + \sum_{i=1}^{8} \alpha_i \epsilon_{t-i}^2$$

$$\sigma_t^2 = 0.0005035 + 0.1241 e_{t-1}^2 + 0.06009 e_{t-2}^2 + 0.0157 e_{t-3}^2 + 0.2195 e_{t-4}^2 + 0.08245 e_{t-5}^2 + 0.11 e_{t-6}^2 + 0.04692 e_{t-7}^2 + 0.05861 e_{t-8}^2$$

Where:

- σ_t^2: Conditional variance of the residuals at time t.
- ω: Constant term representing the baseline variance.
- α_i: Coefficients of lagged squared residuals (ϵ_{t-i}^2).
- ϵ_t: Residuals at time t, assumed to follow a white noise process.

e) Evaluate the unconditional variance of the model's shocks:

$$\text{Unconditional Variance} = \frac{\omega}{1 - \sum_{i=1}^{8} \alpha_i}$$

$$= \frac{5.035 \times 10^{-4}}{1 - (0.1241 + 0.0601 + 0.0157 + 0.2195 + 0.0825 + 0.11 + 0.0469 + 0.0586)}$$

$$= \frac{0.0005035}{1 - 0.7174} = \frac{0.0005035}{0.2826} = 1.781670205 \times 10^{-3} \approx 1.78 \times 10^{-3}$$

∴ , the unconditional (marginal) variance of the shocks in the ARCH(8) model is approximately 1.78×10^{-3}

6) Constructing a 95% one step ahead forecast interval, based on the ARIMA-ARCH model

```
# One-Step-ahead forecast ------------------------------------------------

## ARIMA Forecast Values (from Minitab output):
arima_forecast_2835 = 11.0386                # Forecast for next period
arima_se_2835 = 0.0379696                    # Standard error for the forecast
epsilon_square_2834 = tail(data$Res_Sq, 1)   # Squared Residual of  most recent obs

## ARCH(8) Values (from above):
# omega = params["a0"]
# alpha_params = params[2:9]  # a1 to a8 coefficients
ht_2834 = tail(data$Res_Sq, 1)

## Calculate conditional variance for next period
ht_2835 = omega + sum(alpha_params*epsilon_square_2834)
```

a) The 95% one step forecast interval for my ARIMA-GARCH model:

```
# Calculate the forecast interval (lower and upper bounds)
margin_of_error = 1.96 * sqrt(ht_2835)
forecast_interval_lower = arima_forecast_2835 - margin_of_error
forecast_interval_upper = arima_forecast_2835 + margin_of_error
```

```
## Forecast Interval: [ 10.99461 , 11.08259 ]
```

b) Compared with the 95% one-step forecast interval for ARIMA model from Minitab, earlier:

Forecasts from Time Period 2835

Time Period	Forecast	SE Forecast	95% Limits Lower	95% Limits Upper	Actual
2836	11.0386	0.0379696	10.9641	11.1130	

From the output above, the 95% one-step forecast interval for my ARIMA-GARCH model is {10.99461, 11.08259}, as compared with the ARIMA 95% one-step forecast interval of {10.9641, 11.1130}. We see that the ARIMA-GARCH forecast is narrower and more precise, due to effectively capturing heteroscedasticity, unlike the ARIMA model, which provides broader intervals.

c) The 5th percentile of the conditional distribution of the next period's logged Bitcoin price:

$$5\text{th Percentile} = f_{n+1} + z_{0.05} \cdot \sqrt{h_{n+1}}$$

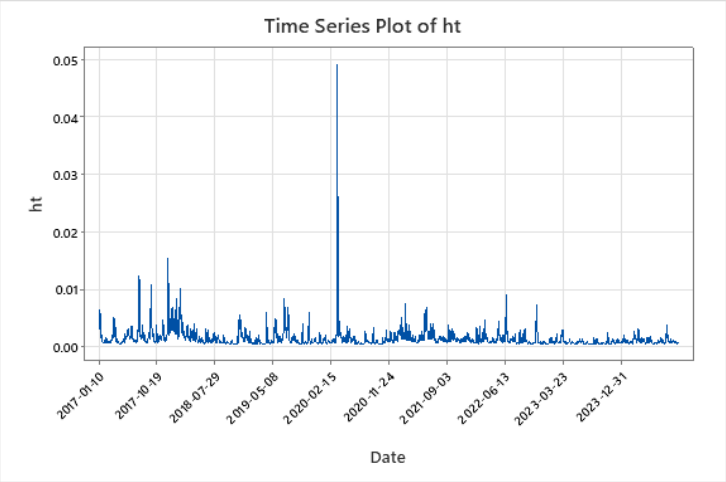
$$5\text{th Percentile} = 11.0386 + (-1.645) \cdot \sqrt{0.000504}$$

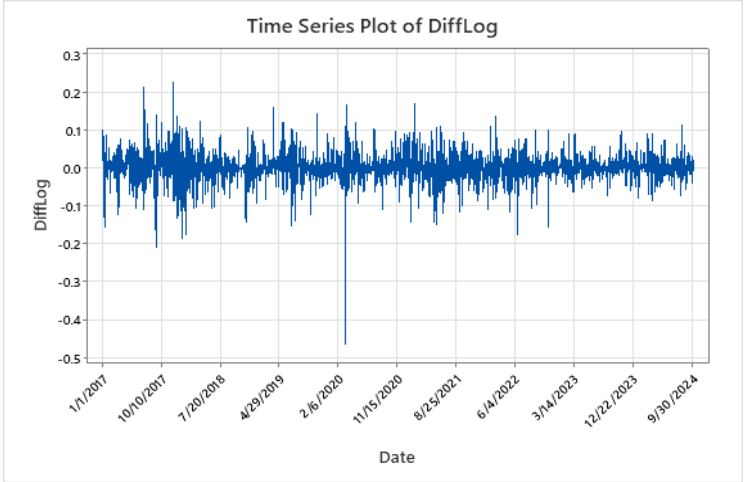
$$5\text{th Percentile} = 11.0386 - 0.03693015841 = 11.00166984 \approx 11.00$$

```
### 5th  percentile of the conditional distribution of next  period -------
# Note: The z-score for the 5th percentile is -1.645
percentile_5th = arima_forecast_2835 + ( -1.645*sqrt(ht_2835) )
cat("5th Percentile of the Forecast: ", percentile_5th, "\n")
```

```
## 5th Percentile of the Forecast:  11.00168
```

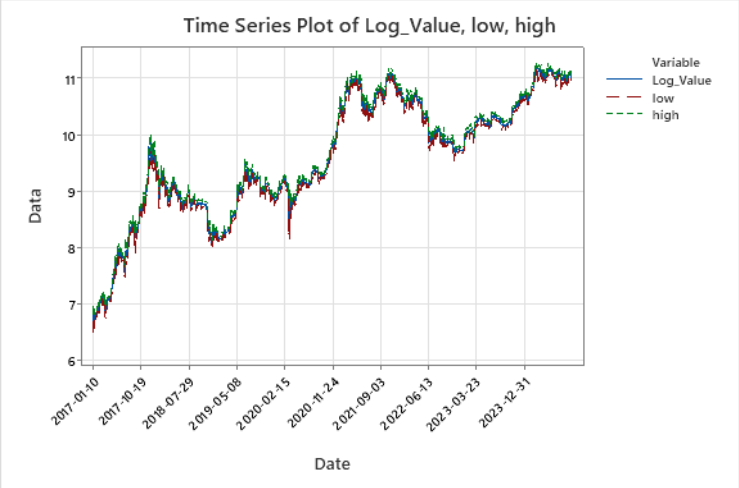
7) Plot of the Unconditional Variances (h_t) — for the fitted ARCH model





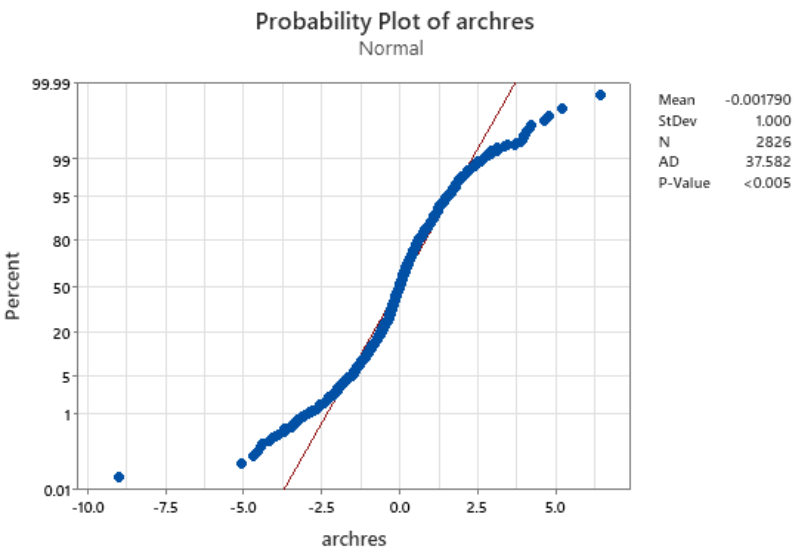
From the plots above, we see that bursts of high volatility in the conditional variances (h_t) reflects a sharp increase in volatility (magnitude), while the corresponding downward spike in the differenced log values indicates a significant negative movement in Bitcoin prices. This agreement highlights that the ARCH model effectively captures periods of heightened volatility, even when the direction of price changes differs.

8) Time series plot of historical data together with the ARIMA-ARCH one-step ahead 95% forecast intervals



The ARIMA-ARCH one-step ahead 95% forecast intervals align well with the observed data, showing good accuracy and practical usefulness. The intervals adapt to volatility, narrowing in stable periods and widening during high volatility, such as in early 2020. However, a potential concern with this model is the risk of overfitting. It may be too effective at finding intervals for specific observations, which could limit its ability to generalize and accurately forecast future values.

9) Time series plot of historical data together with the ARIMA-ARCH one-step ahead 95% forecast intervals



From the normal probability plot of the ARCH residuals (archres), we observe significant deviations from the straight line, particularly in the tails.

This indicates that the residuals do not follow a normal distribution and instead exhibit leptokurtosis ("long-tailedness").

Thus, the model does not appear to have fully captured the heavy tails in the data. This suggests that the ARCH model could potentially be improved by considering alternative specifications, such as a GARCH model or assuming a non-normal error distribution (e.g., t-distribution).

10) Prediction failures

Failures are the number of times the predicted range (95% prediction interval) didn't include the actual value. This happens when the model's error (residual) is too large, beyond 1.96 in either direction.

C10
failures
159

- Number of failures = 159
- Number of predictions = 2826 (after dropping missing values from residual analysis, etc.)
- Percentage of the time the intervals failed = $\frac{159}{2826} \times = 5.63\%$

From the above, we can conclude that the percentage of the time the intervals failed is about 5%, which aligns with the expected failure rate for a 95% prediction interval.