IE313: Time Series Analysis
Problem Sets 3

Student Name: Nguyen Minh Duc
Student ID: 20202026

Problem 1. Identify the following as specific ARIMA models. That is, what are p, d, and q and what are the values of the parameters (the ϕ's and θ's)? And, please answer whether they are stationary or not.

(a) $Y_t = Y_{t-1} - 0.25Y_{t-2} + e_t - 0.1e_{t-1}$.

(b) $Y_t = 2Y_{t-1} - Y_{t-2} + e_t$.

(c) $Y_t = 0.5Y_{t-1} - 0.5Y_{t-2} + e_t - 0.5e_{t-1} + 0.25e_{t-2}$.

Solution.

(a) The series has the form of the ARMA$(2,1)$ model where the parameters are $\phi_1 = 1$, $\phi_2 = -0.25$, and $\theta_1 = 0.1$. Let's take a loot at its characteristic equation:

$$\begin{aligned} 1 - \phi_1 x - \phi_2 x^2 &= 0 \\ 1 - x + 0.25x^2 &= 0 \\ 4 - 4x + x^2 &= 0 \\ (2-x)^2 &= 0 \\ x &= 2 \\ \implies |x| &> 1. \end{aligned}$$

Thus, $\{Y_t\}$ is stationary, which implies $\{Y_t\}$ is an ARIMA$(2,0,1)$ model.

(b) The series has the form of the AR(2) model where the parameters are $\phi_1 = 2$, and $\phi_2 = -1$. Let's take a loot at its characteristic equation:

$$\begin{aligned} 1 - \phi_1 x - \phi_2 x^2 &= 0 \\ 1 - 2x + x^2 &= 0 \\ (1-x)^2 &= 0 \\ x &= 1 \\ \implies |x| &\leq 1. \end{aligned}$$

Thus, $\{Y_t\}$ is not stationary. Taking difference of $\{Y_t\}$, we have

$$\nabla Y_t = Y_t - Y_{t-1} = Y_{t-1} - Y_{t-2} + e_t,$$

which is an AR(2) model with parameters $\phi_1 = 1$ and $\phi_2 = -1$. Solving its characteristic equation yields

$$
\begin{aligned}
1 - \phi_1 x - \phi_2 x^2 &= 0 \\
1 - x + x^2 &= 0 \\
\left(\frac{1}{2} - x\right)^2 + \frac{3}{4} &= 0 \\
\left(\frac{1}{2} - x\right)^2 &= -\frac{3}{4} \\
\frac{1}{2} - x &= \pm\frac{i\sqrt{3}}{2} \\
x &= \frac{1}{2} \pm \frac{i\sqrt{3}}{2} \\
\implies |x| &= \sqrt{\frac{1}{4} + \frac{3}{4}} = 1 \leq 1,
\end{aligned}
$$

which implies that ∇Y_t is still not stationary. Taking the difference one more time, we have

$$\nabla^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2} = e_t,$$

which is an MA(0) process. Therefore, $\nabla^2 Y_t$ is stationary, which implies $\{Y_t\}$ is an ARIMA$(0, 2, 0)$ model.

(c) The series has the form of the ARMA$(2, 2)$ model where the parameters are $\phi_1 = 0.5$, $\phi_2 = -0.5$, $\theta_1 = 0.5$, and $\theta_2 = -0.25$. Let's take a loot at its characteristic equation:

$$
\begin{aligned}
1 - \phi_1 x - \phi_2 x^2 &= 0 \\
1 - 0.5x + 0.5^2 &= 0.
\end{aligned}
$$

Note that $\begin{cases} \phi_2 + \phi_1 = -0.5 + 0.5 = 0 < 1 \\ \phi_2 + \phi_1 = -0.5 - 0.5 = -1 < 1 \\ |\phi_2| = 0.5 < 1 \end{cases}$, which implies that $|x| > 1$, hence $\{Y_t\}$ is stationary.
Therefore, $\{Y_t\}$ is an ARIMA$(2, 0, 2)$ process.

Problem 2.

(a) Download the given data `Samsung_close.csv` and perform exploratory data analysis. Then try to transform the data into a stationary series.

(b) Find at least two candidate orders for the ARMA (or ARIMA) model. You should discuss why. Discuss whether the estimated trend fits your data or not.

(c) Perform residual analysis. Discuss the results.

Solution.

(a) The provided data contains the closing stock price of Samsung from `2021-01-04` to `2022-10-12`. Table 1 contains the first 5 days in the data set. Plotting the data results in a time series in Figure 2.

Date	Close
2021-01-04	83000
2021-01-05	83900
2021-01-06	82200
2021-01-07	82900
2021-01-08	88800

Table 1: Data preview

However, before getting into more details, let's check if there are any missing dates in the data set, please refer to Figure 1 for more details. As we can see, there are 209 missing dates scattered all over the date range. This could be that the stock market closes on holidays and weekends, hence, the missing dates. Plotting the data results in Figure 2. As we can see, there is a strong decreasing trend in the data and some underlying weak seasonality.

```
check_missing_date = df.reindex(pd.date_range(start = '2021-01-04', end = '2022-10-12' )).isnull().all(1)
check_missing_date[check_missing_date == True]
```

[138] ✓ 0.0s

```
2021-01-09    True
2021-01-10    True
2021-01-16    True
2021-01-17    True
2021-01-23    True
              ...
2022-10-02    True
2022-10-03    True
2022-10-08    True
2022-10-09    True
2022-10-10    True
Length: 209, dtype: bool
```

Figure 1: Missing dates

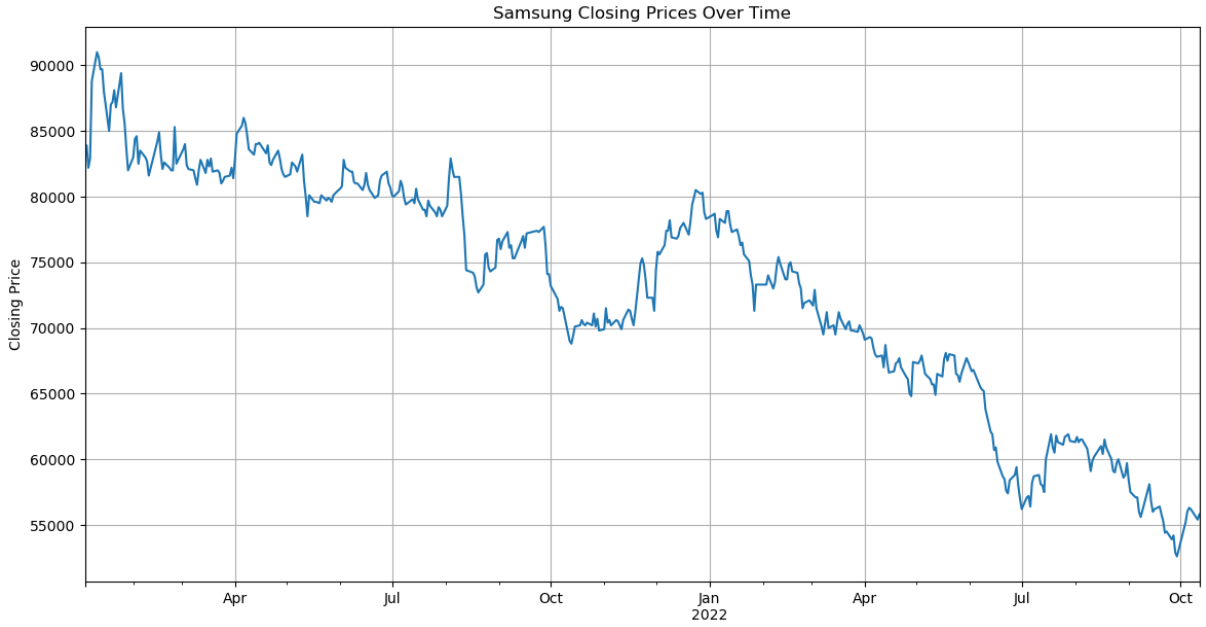

Figure 2: Time Series Plot of Samsung's stock

Now, let's check the stationarity of our series. Figure 3 shows plots for ACF, PACF, and QQ plots. As we can observe, there are large correlations among neighboring points in our data set and the PACF spikes at lag 1 and tails off at later lags, indicating that there may be an AR(1) component in our series. Moreover, in the QQ plots, the values follow a straight line quite close, but there are still a lot of outliers. Thus, it is, with high probability, that the time series is not stationary, hence, using an ARMA model directly would be inappropriate. To efficiently analyze Samsung's stock prices, we need to transform the data to a stationary one. To begin with, the values of the stock price seem large, we can try to make it smaller and also remove some potential exponential components by using log transformation. Figure 4 contains the visualizations for the log transformation and two differences.

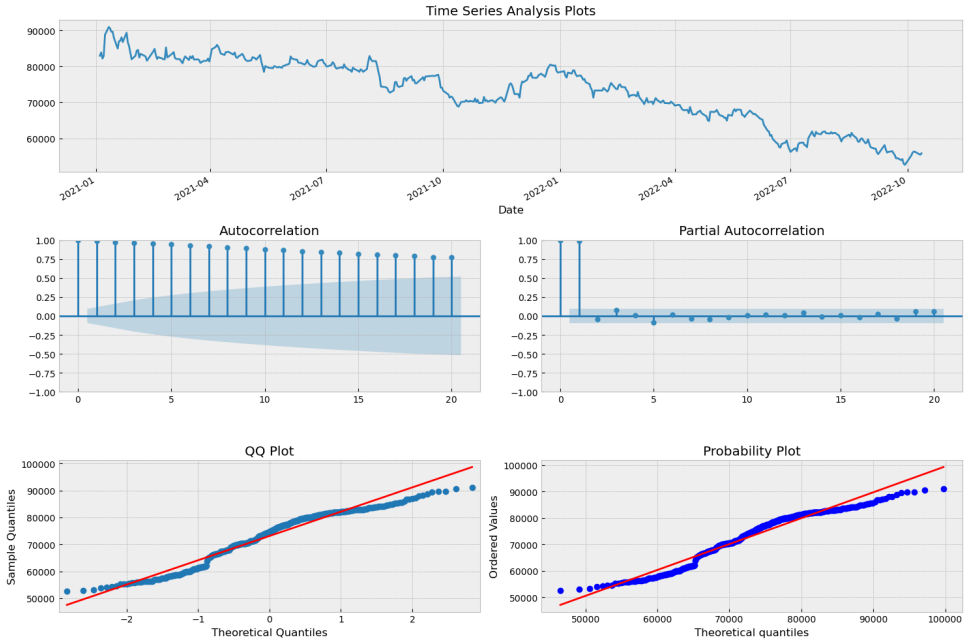

Figure 3: Correlations and QQ plots

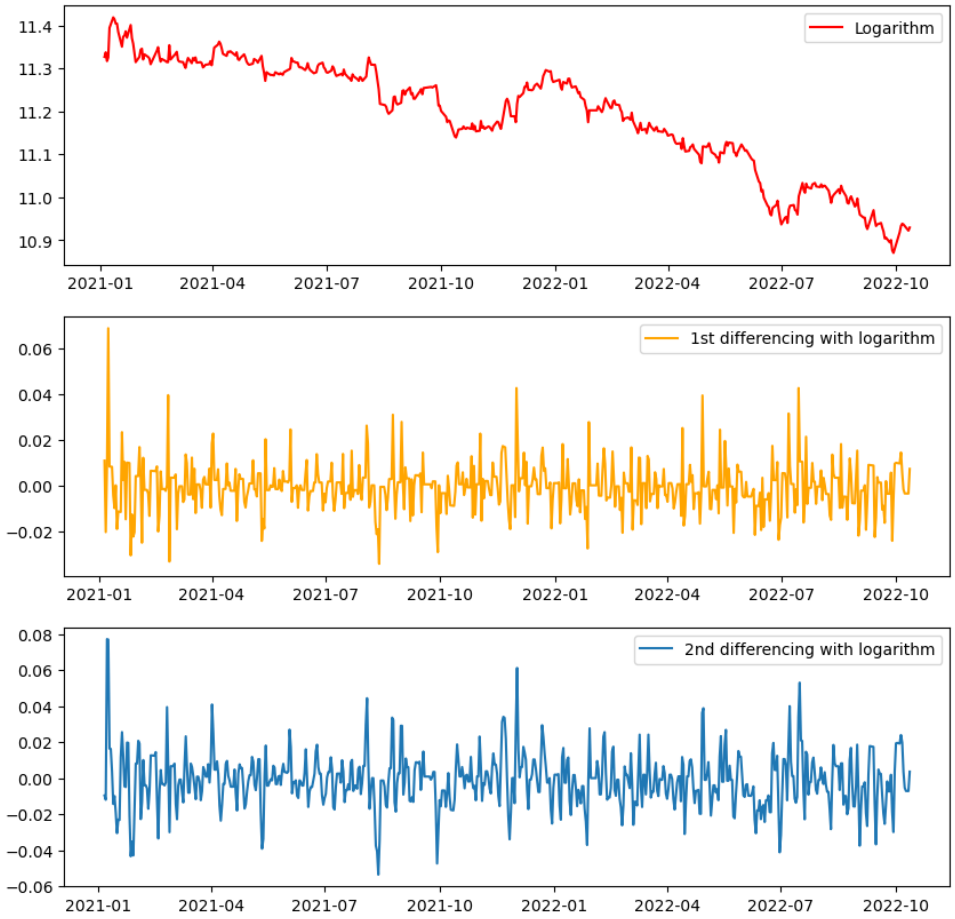

Figure 4: Log Transformation with Differences

Both differences look stationary. To be sure, let's conduct an Augmented Dickey-Fuller test to see if the stationarity of the differences is statistically significant.

```
---1st differencing with logarithm---
ADF Statistic: -23.262855
Critical Values @ 0.05: -2.87
p-value: 0.00000000000000000000

---2nd differencing with logarithm---
ADF Statistic: -6.482134
Critical Values @ 0.05: -2.87
p-value: 0.00000001287202098441
```

Looking at the result of the ADF test, we can confirm that the fact "the first and the second difference with logarithm" is statistically significant. For the sake of parsimony, let's consider the first difference with logarithm. Looking at the Time Series Analysis Plots (Figure 5) further ensures the stationarity of our transformation. Most values ACF and PACF are well within the 2 standard deviation zone and the QQ plot follows the straight line quite closely.

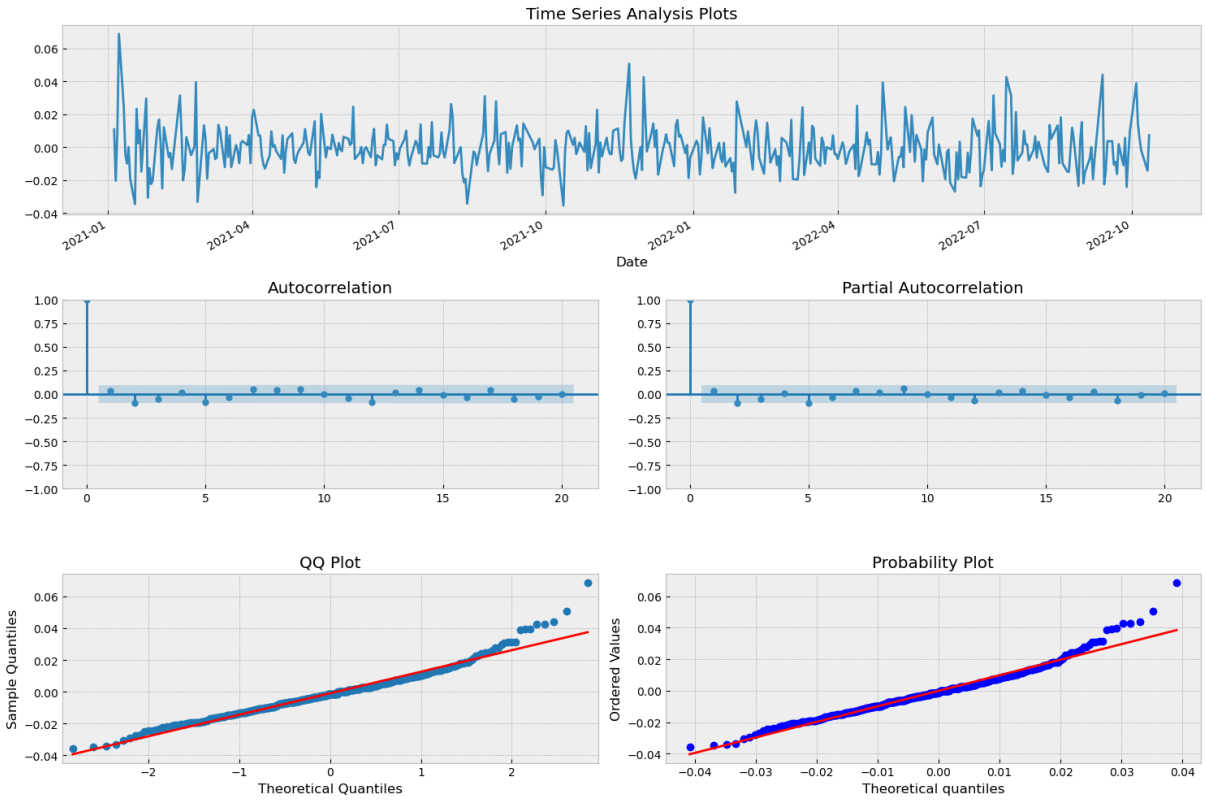

Figure 5: Time Series Analysis Plots

(b) Let's find ARIMA models that fit well with our data. Note that since we have already taken the difference once, the d order of the ARIMA model should be at least 1. Looking at the partial autocorrelation function (PACF) plot, it looks like a "sine wave". Most of them are in the blue boundary. It suggests $p = 0$, which is the order of the AR(p) model. However, when looking at the autocorrelation function (ACF) plot, they are also well within the blue boundary, hence, there is a clear cut-off lag at $q = 0$, which is the order of MA(q) model. Therefore, one promising candidate would be ARIMA$(0, 1, 0)$. Another model worthy of considering is ARIMA$(1, 1, 0)$. This model appears a lot in financial settings, which is what our data set is based on. Moreover, looking at the original plot (Figure 3), we can see that the model has a significantly large PACF at lag 1, suggesting an AR(1) component. Please refer to Figure 6 and 7 for the estimation results of the two models.

SARIMAX Results

Dep. Variable:	value	No. Observations:	438
Model:	ARIMA(0, 1, 0)	Log Likelihood	1259.428
Date:	Mon, 30 Oct 2023	AIC	-2516.857
Time:	14:15:06	BIC	-2512.777
Sample:	0	HQIC	-2515.247
	- 438		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	0.0002	8.7e-06	21.113	0.000	0.000	0.000

Ljung-Box (L1) (Q):	0.60	Jarque-Bera (JB):	142.56
Prob(Q):	0.44	Prob(JB):	0.00
Heteroskedasticity (H):	1.04	Skew:	0.79
Prob(H) (two-sided):	0.80	Kurtosis:	5.31

Figure 6: Estimation Result for ARIMA$(0, 1, 0)$

The result has a low AIC score, indicating a good fit for our data. Also, the constant covariance is statistically significant as the p-value is less than 0.05. Also, the moderate Ljung-Box statistic suggests that there is little autocorrelation in the residuals.

SARIMAX Results

Dep. Variable:	value	No. Observations:	438
Model:	ARIMA(1, 1, 0)	Log Likelihood	1259.805
Date:	Mon, 30 Oct 2023	AIC	-2515.610
Time:	14:17:41	BIC	-2507.450
Sample:	0	HQIC	-2512.390
	- 438		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0415	0.041	1.015	0.310	-0.039	0.122
sigma2	0.0002	8.69e-06	21.106	0.000	0.000	0.000

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	140.15
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	1.04	Skew:	0.78
Prob(H) (two-sided):	0.83	Kurtosis:	5.29

Figure 7: Estimation Result for ARIMA$(1, 1, 0)$

The result also has a low AIC score, indicating a good fit for our data. Also, the constant variance is not quite statistically significant as the p-value is greater than 0.05. Also, the high Ljung-Box statistic suggests that there is almost no autocorrelation in the residuals.

(c) Let's conduct some residual analysis.

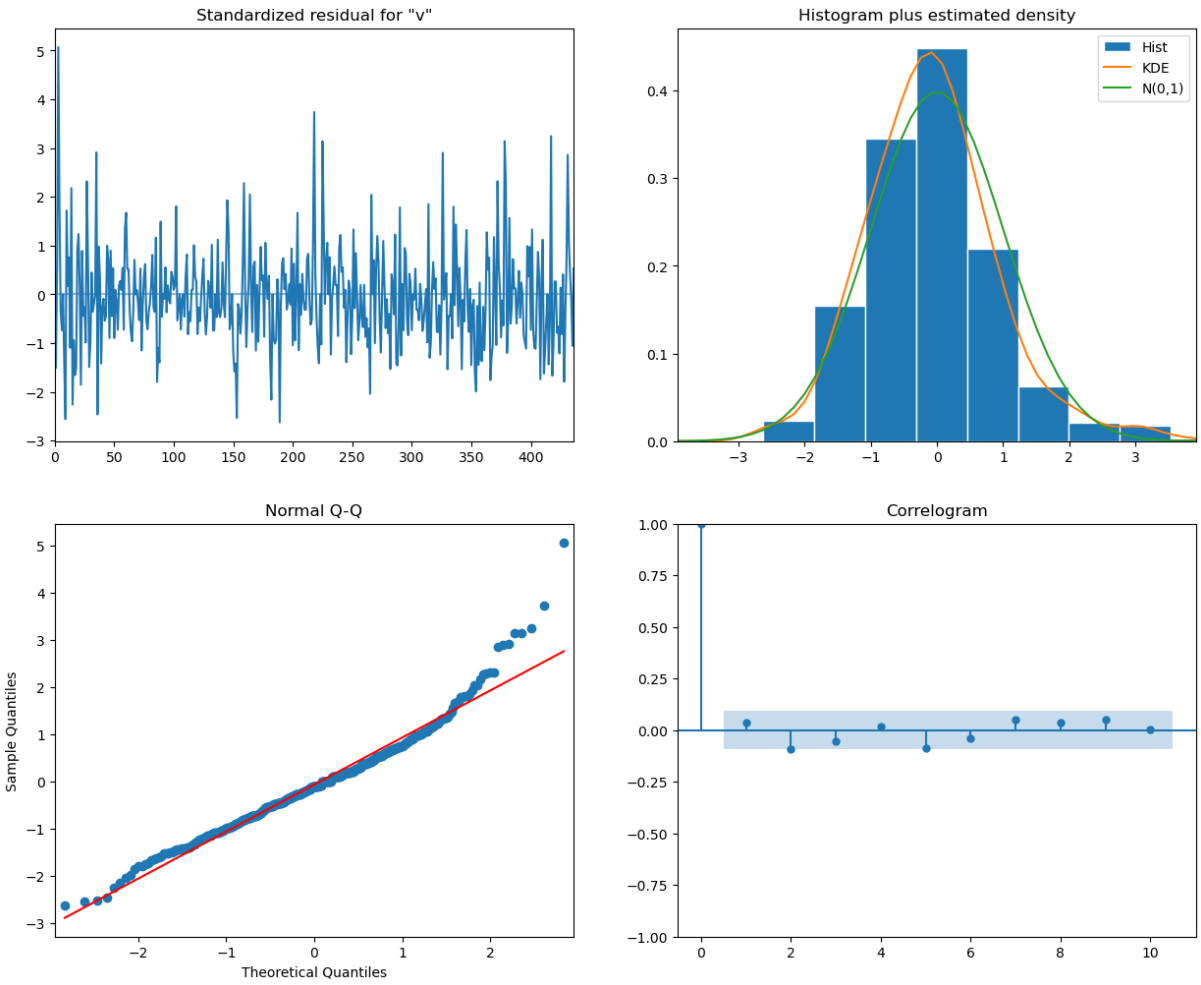

Figure 8: Residual Analysis of ARIMA$(0, 1, 0)$

Please refer to Figure 8 for the plots in residual analysis of ARIMA$(0, 1, 0)$. The plot for standardized residual looks stationary, the residuals are generally centered around zero, and there is no clear trend. However, there may be some recurring patterns left in the residuals. As for the autocorrelations, one could easily observe that all of the lags except for lag 0 have their autocorrelation well within the two-standard-error zone, which is the expected behavior of a white noise process, hence, the model has eliminated most of the autocorrelation in the pattern of the residuals. Looking at the histogram of the residuals, it resembles the bell curve of the normal distribution and has a slight right-skewness. Also, it does have a much higher kurtosis. As for the Q-Q plot, most of the residuals are on the critical $y = x$ line, but some of them are off at the tails, especially the right tail. To sum up the residual analysis, the residuals of our model look stationary indicating that our suggested model fits our data quite well.

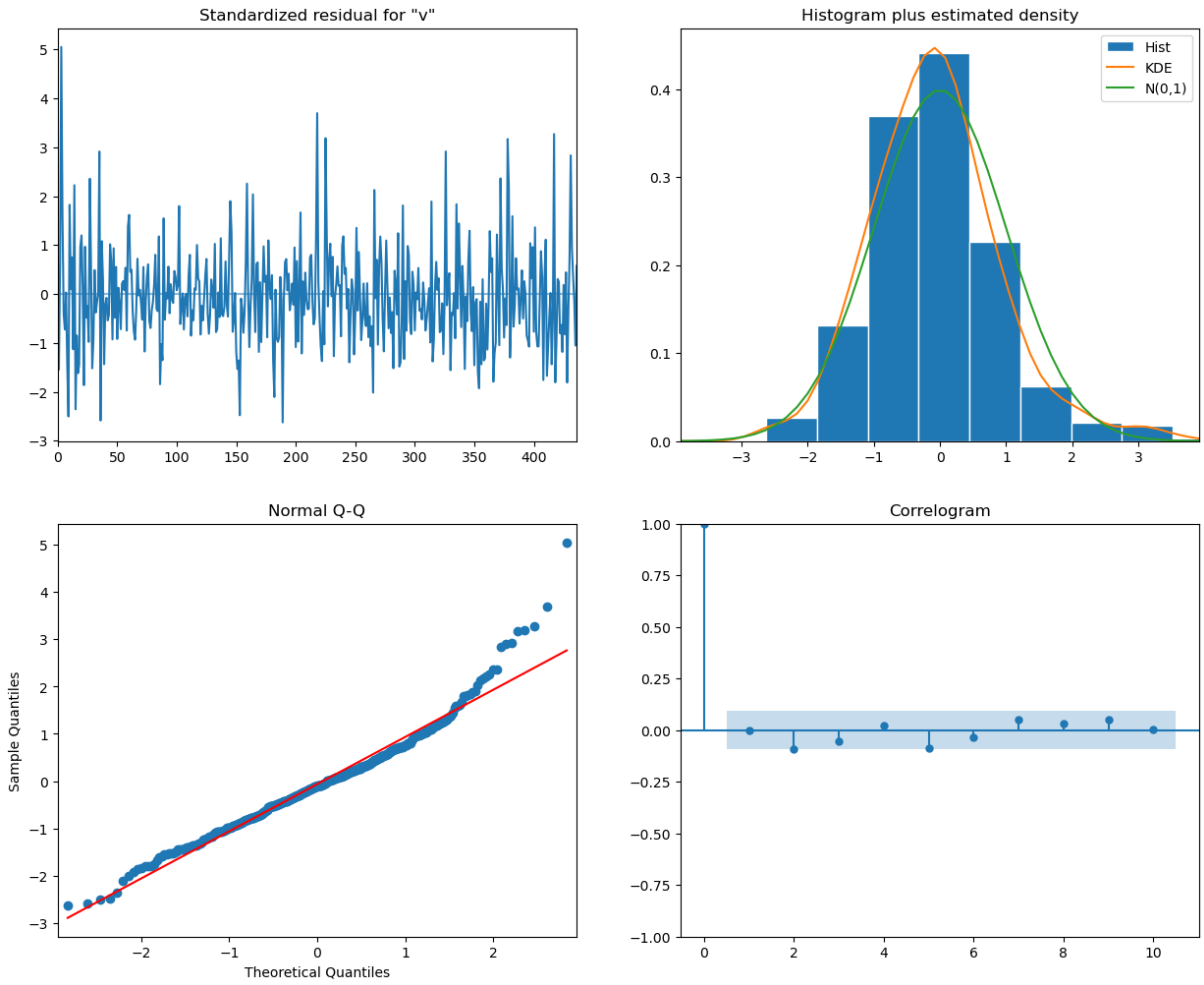

Figure 9: Residual Analysis of ARIMA$(1,1,0)$

Please refer to Figure 9 for the plots in residual analysis of ARIMA$(1,1,0)$. The plot for standardized residual looks stationary, the residuals are generally centered around zero, and there is no clear trend. Looking at the histogram of the residuals, it resembles the bell curve of the normal distribution and has a slight right-skewness. Also, it does have a much higher kurtosis. As for the Q-Q plot, most of the residuals are on the critical $y = x$ line, but some of them are off at the tails, especially the right tail where the outliers are off to a high degree. As for the autocorrelations, one could easily observe that all of the lags except for lag 0 have their autocorrelation well within the two-standard-error zone, which is the expected behavior of a white noise process, hence, the model has eliminated most of the autocorrelation in the pattern of the residuals. To sum up the residual analysis, the residuals of our model look stationary indicating that our suggested model fits our data quite well. To fit better, we need to consider those outliers that are messing up with the normality of our residuals.

Problem 3. Consider a stationary process $\{Y_t\}$. Show that if $\rho_1 < 0.5$, ∇Y_t has a larger variance than does Y_t.

Since $\{Y_t\}$ is stationary, $\text{Var}(Y_t) = \gamma_0$, and $\rho_1 = \dfrac{\text{Cov}\,(Y_t, Y_{t-1})}{\sqrt{\text{Var}\,(Y_t)\,\text{Var}\,(Y_{t-1})}} = \dfrac{\gamma_1}{\sqrt{\gamma_0\gamma_0}} = \dfrac{\gamma_1}{\sqrt{{\gamma_0}^2}} = \dfrac{\gamma_1}{\gamma_0} < \dfrac{1}{2}$
$\implies \gamma_0 - 2\gamma_1 > 0$..

Note that $\text{Var}\,(\nabla Y_t) = \text{Var}\,(Y_t - Y_{t-1}) = \text{Var}\,(Y_t) + \text{Var}\,(Y_{t-1}) - 2\,\text{Cov}\,(Y_t, Y_{t-1}) = \gamma_0 + \gamma_0 - 2\gamma_1$.

Since $\gamma_0 - 2\gamma_1 > 0$, $\text{Var}\,(\nabla Y_t) > \gamma_0 = \text{Var}\,(Y_t)$. Therefore, ∇Y_t has a larger variance than does Y_t.