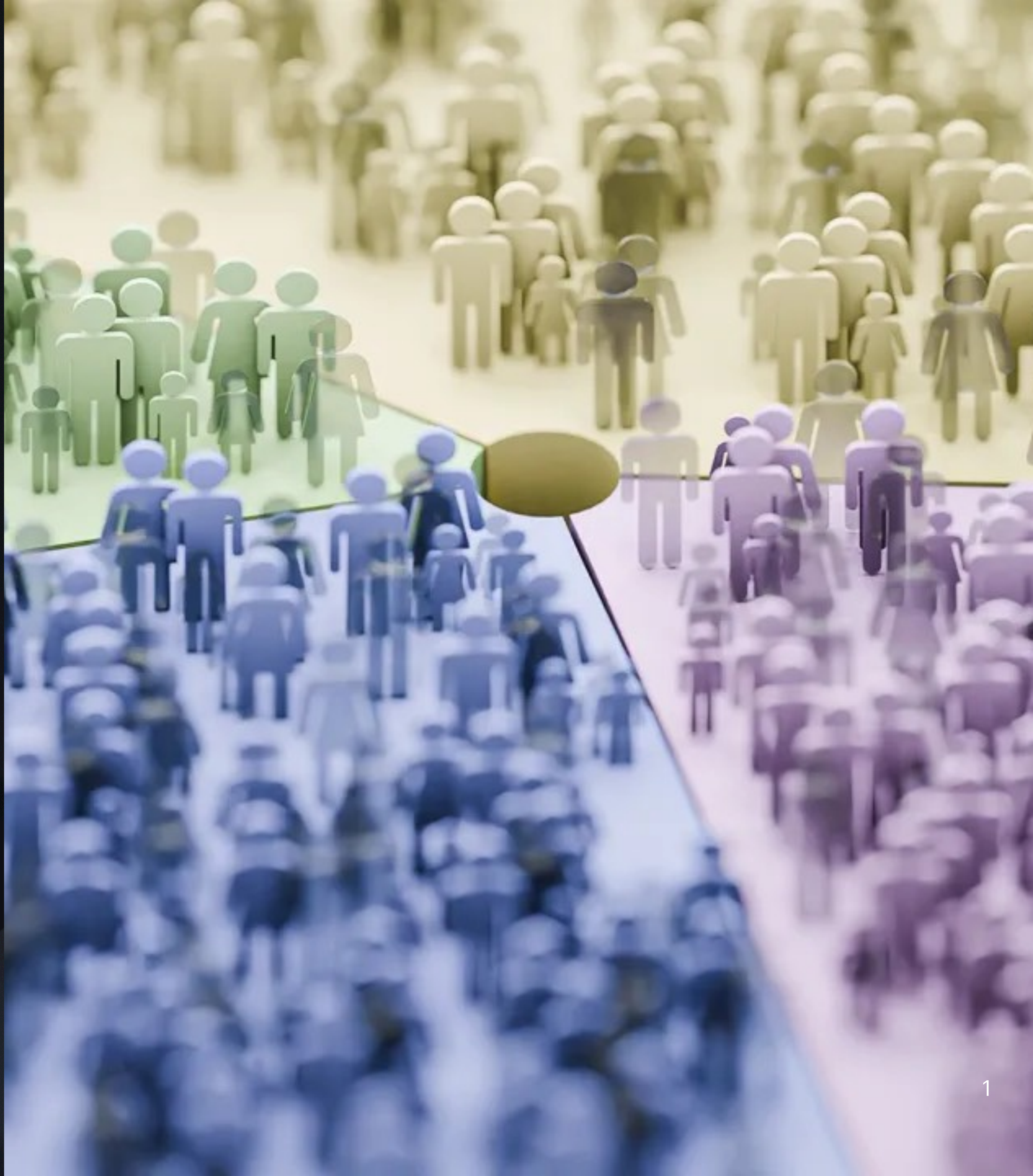
Forecasting HIV Trends in Singapore (1985 - 2023) using ARIMA Modelling

Machine Learning for Biomedical Informatics



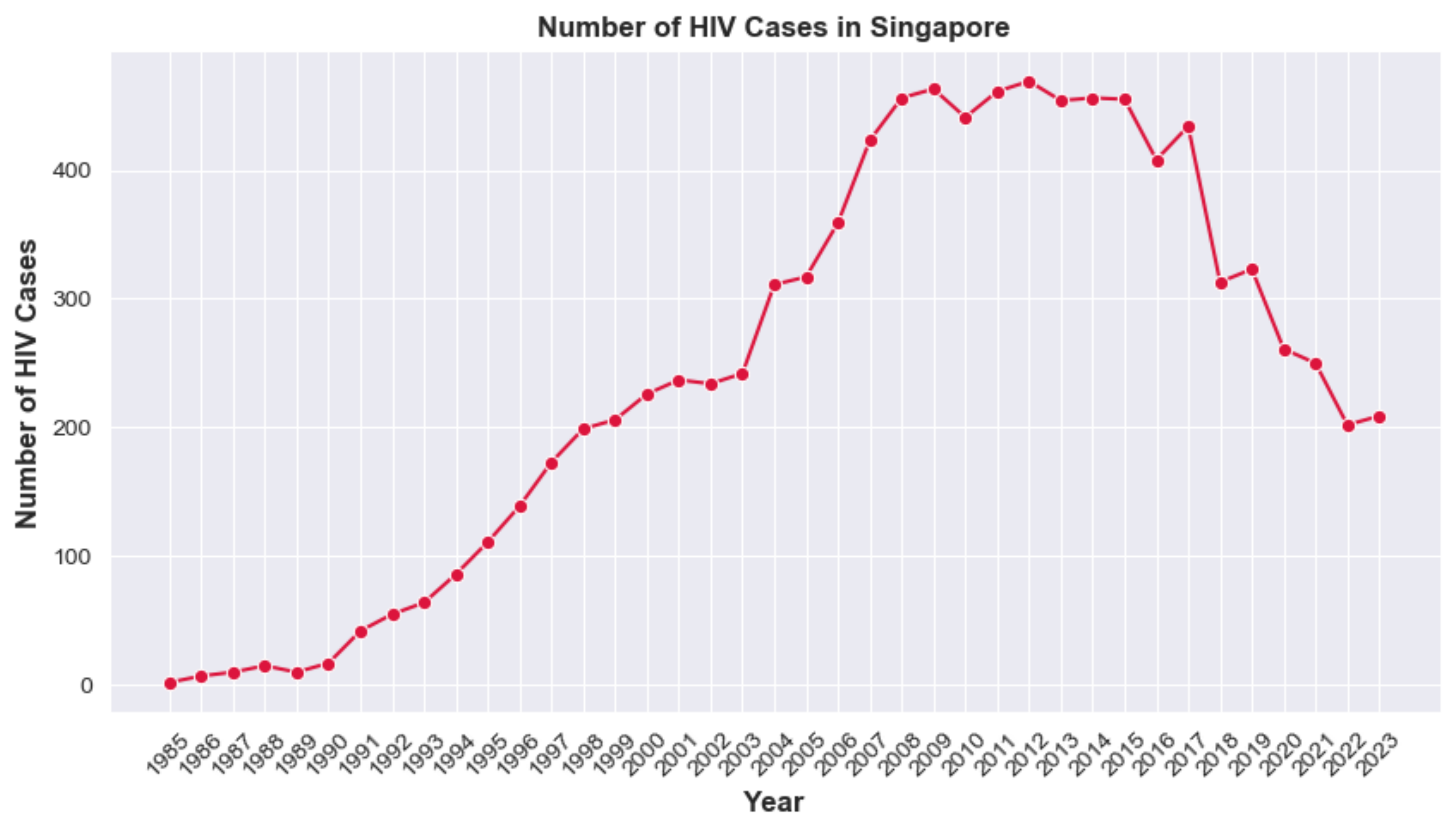
Objectives

- To pre-process and analyze historical HIV case data in Singapore from 1985 to 2023.
- To determine the optimal ARIMA model parameters (p, d, q) for accurately capturing the trends in HIV cases.
- To validate the model using performance metrics such as AIC or BIC for reliability.
- To forecast HIV case trends for the coming years aiding public health initiatives and long-term strategic planning.

HIV Cases in Singapore

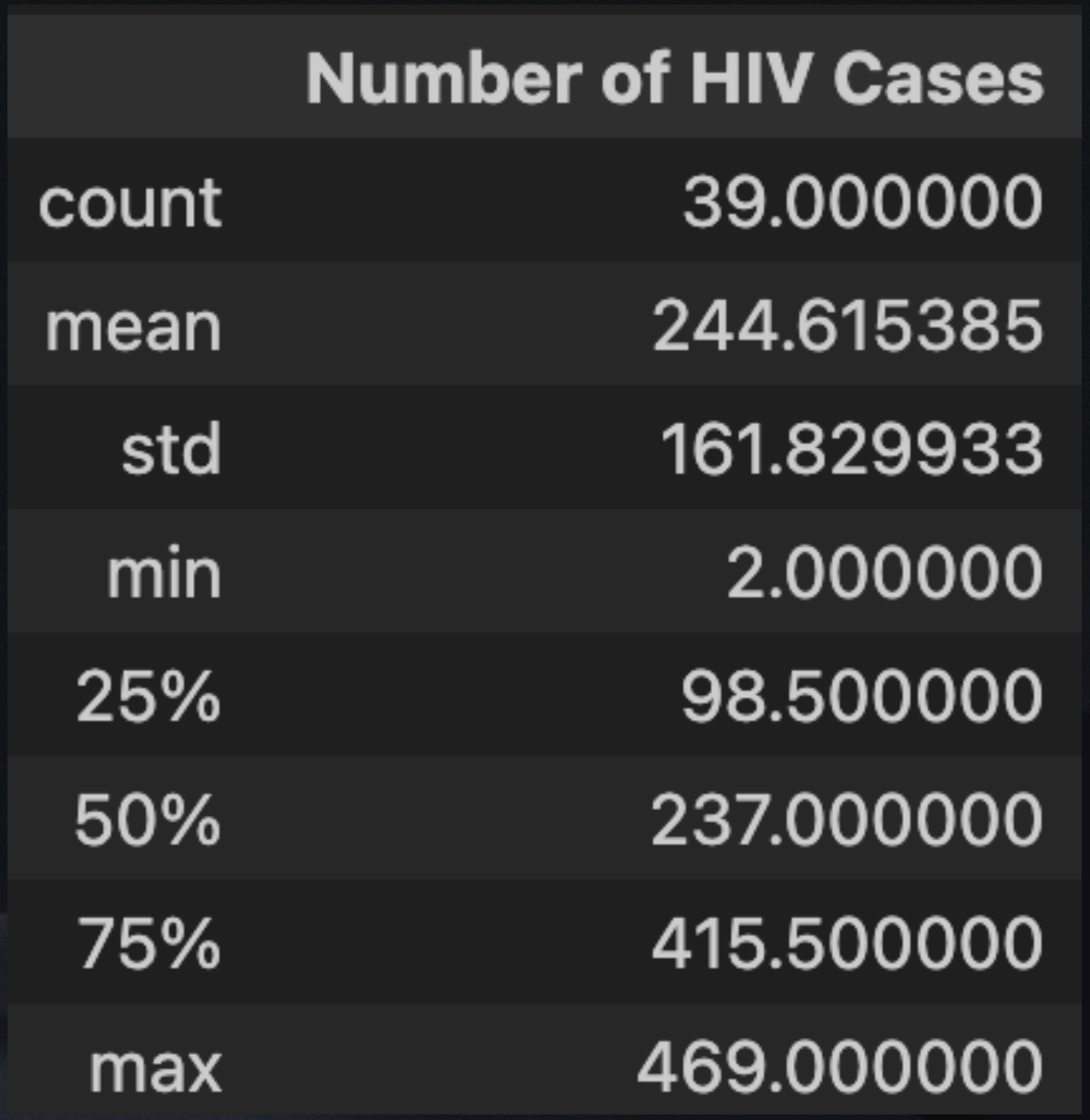
Trends

- The number of HIV cases rose steadily over two decades which reflects increased reporting, awareness, and possible transmission rates.
- The cases plateaued at their highest trends which indicates a peak in reported cases during these years.
- A significant decline in cases is observed post-2013, - potentially due to effective public health interventions, awareness campaigns, and advancements in medical treatments.



HIV Cases in Singapore

Summary Statistics



	Number of HIV Cases
count	39.000000
mean	244.615385
std	161.829933
min	2.000000
25%	98.500000
50%	237.000000
75%	415.500000
max	469.000000

- **Count:** There are 39 data points (observations) for the number of HIV cases.
- **Mean:** The average number of HIV cases across all observations is 244.62.
- **Std:** The variation in the number of HIV cases is quite high with SD of 161.83.
- **Min:** The smallest number of HIV cases observed is 2.
- **25% (Q1):** 25% of the data points have fewer than 98.5 HIV cases.
- **50% (Q2):** The median number of HIV cases is 237 meaning that 50% of the data points are below and above of this value.
- **75% (Q3):** 75% of the data points have fewer than 415.5 HIV cases.
- **Max:** The largest number of HIV cases observed is 469.

Initial Modelling

Differencing

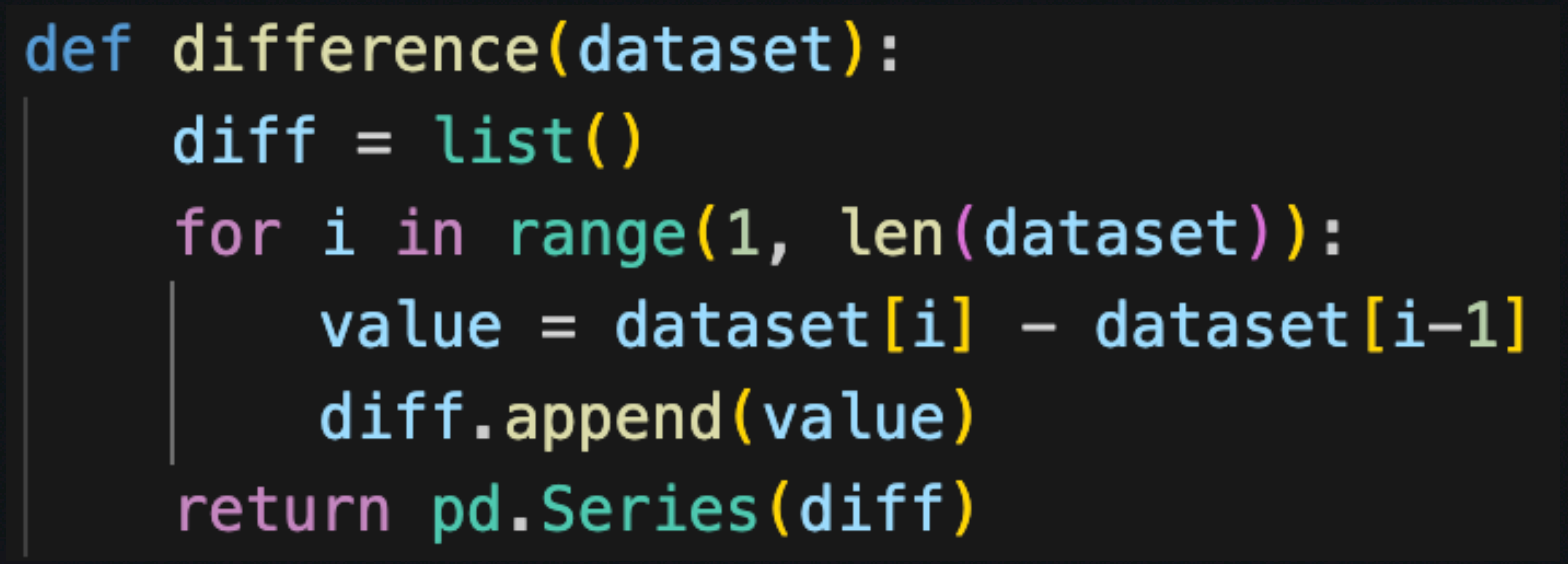
- **Primary Purpose:** To remove trends or seasonality in time series data - to make the data stationary (constant mean and variance over time).

- **How it works?**

1. Iterates through the dataset, starting from the second element.

2. For each element, it calculates the difference between the current value and the previous value.

3. These differences are stored in a new list, which returned as a pandas Series.

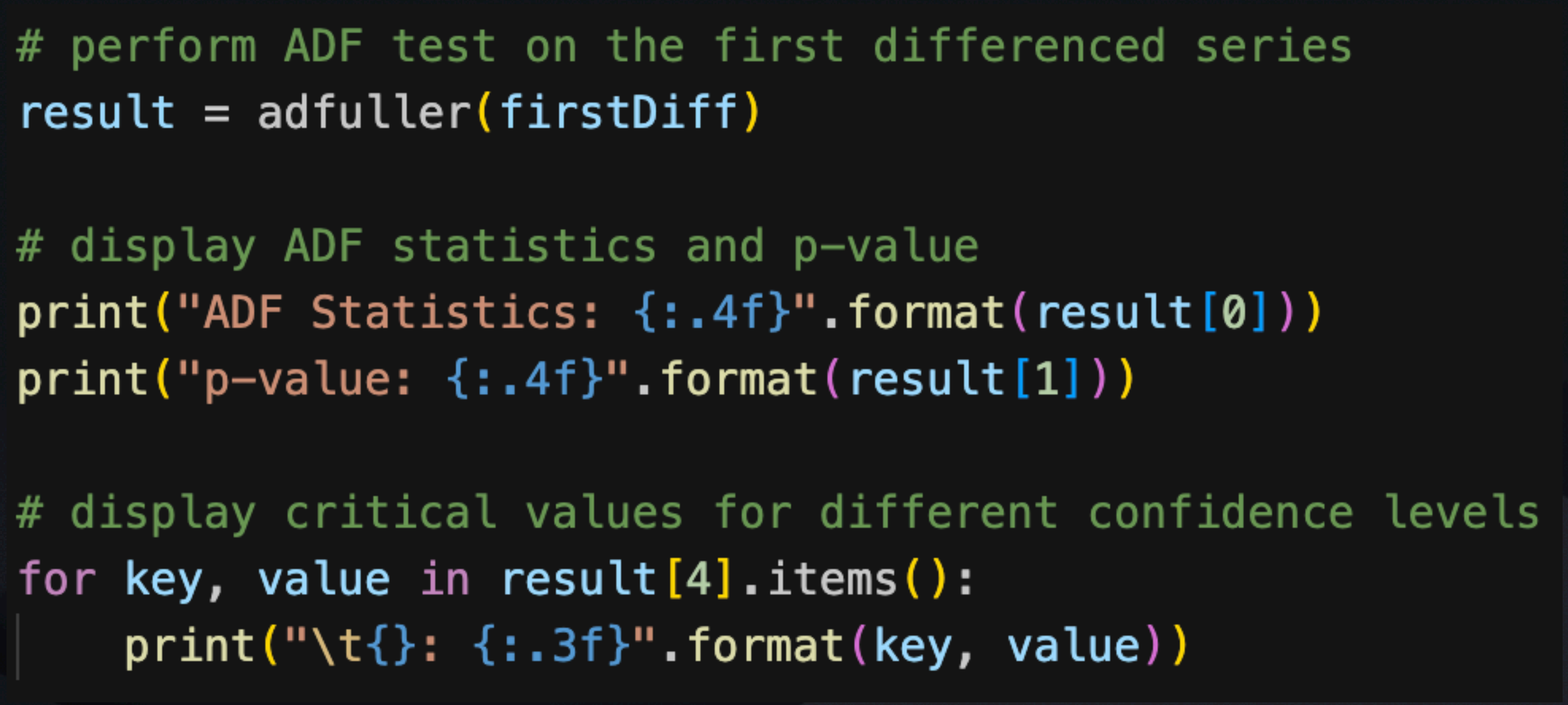


```python
def difference(dataset):
    diff = list()
    for i in range(1, len(dataset)):
        value = dataset[i] - dataset[i-1]
        diff.append(value)
    return pd.Series(diff)
```

Initial Modelling

Augmented Dickey-Fuller (ADF)

- **Primary Purpose:** To determine whether a time series is stationary or a trend.



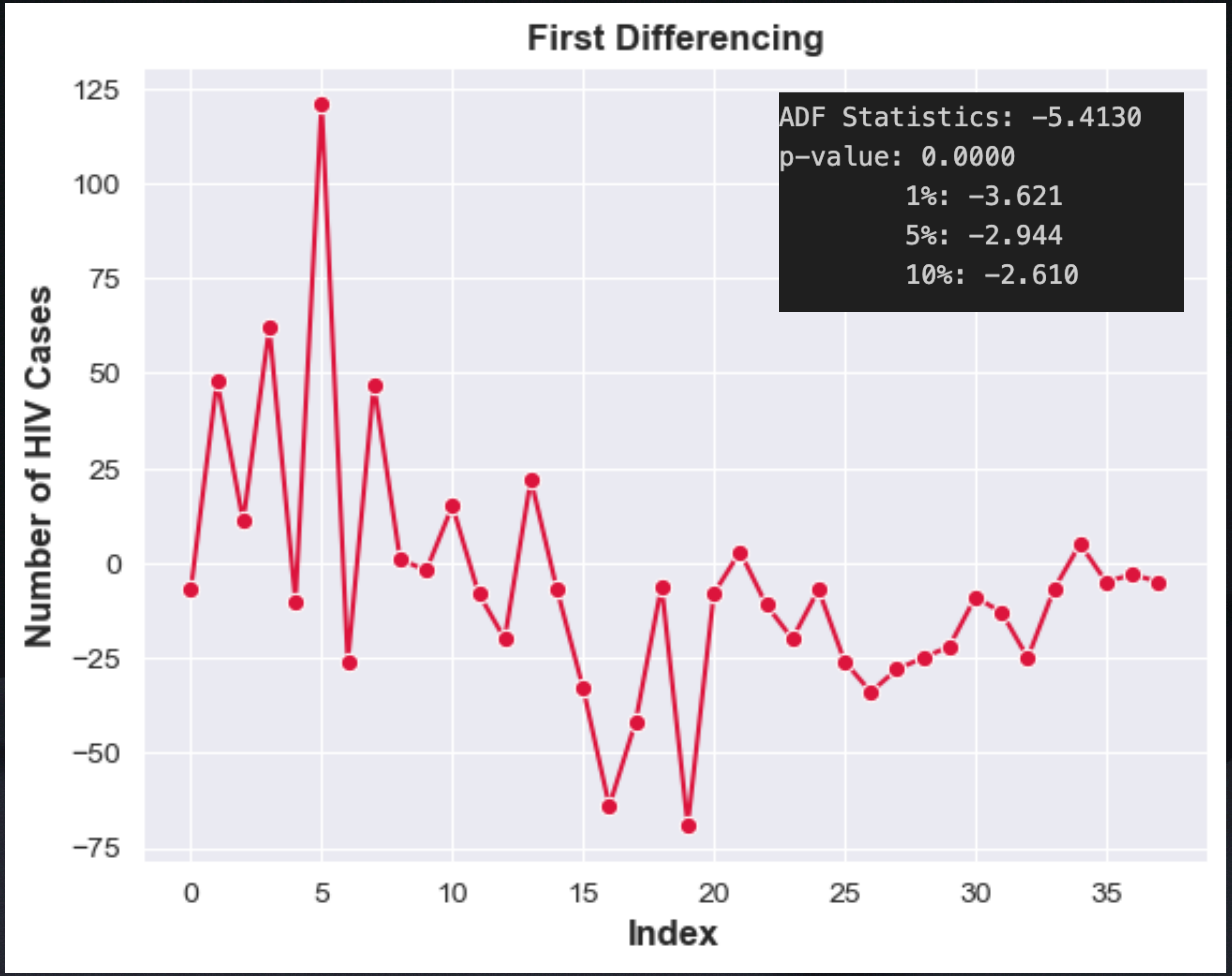
```python
# perform ADF test on the first differenced series
result = adfuller(firstDiff)

# display ADF statistics and p-value
print("ADF Statistics: {:.4f}".format(result[0]))
print("p-value: {:.4f}".format(result[1]))

# display critical values for different confidence levels
for key, value in result[4].items():
    print("\t{}: {:.3f}".format(key, value))
```

Initial Modelling

First-order Differencing



```
ADF Statistics: -5.4130
p-value: 0.0000
        1%: -3.621
        5%: -2.944
        10%: -2.610
```

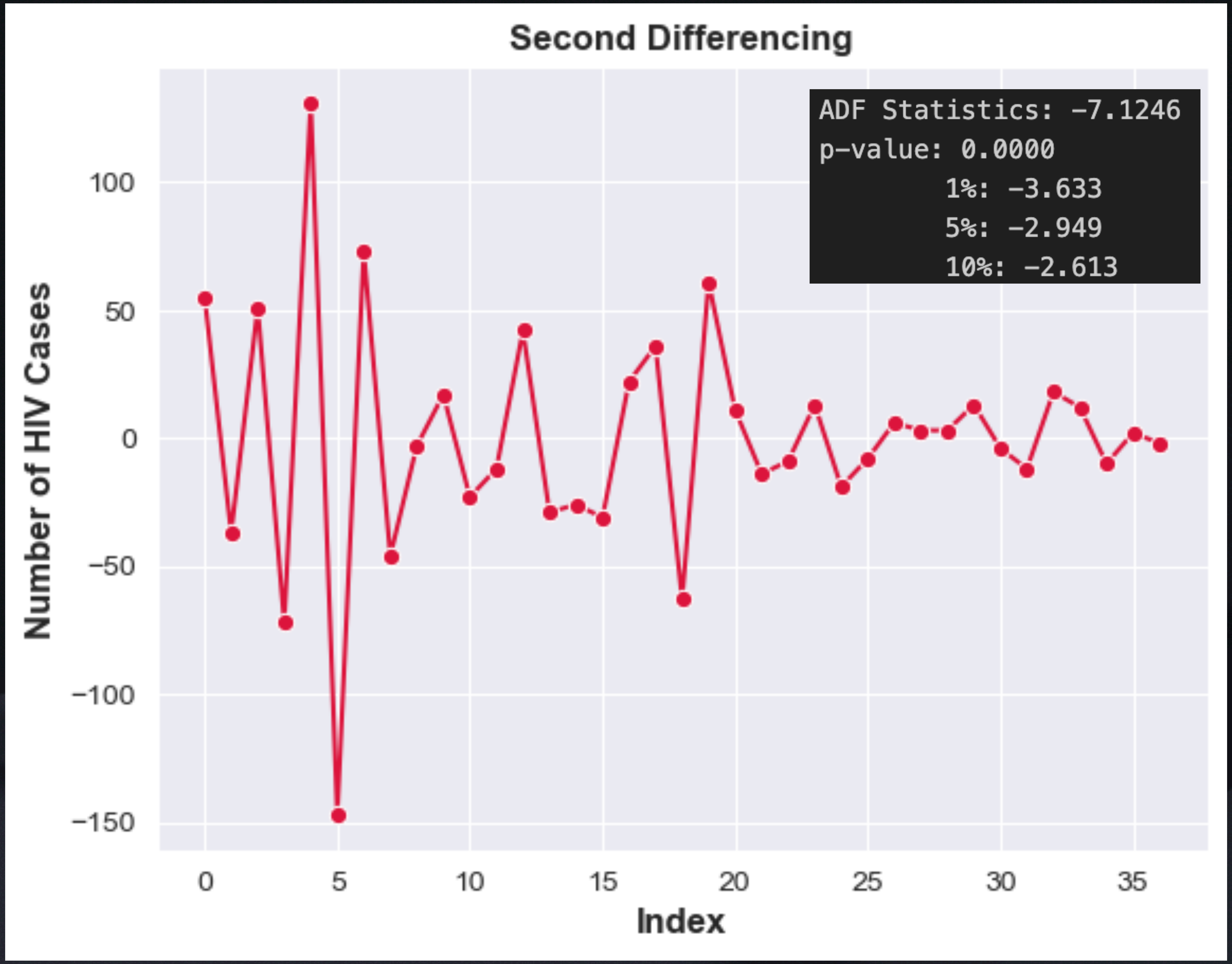
- Since the p-value is below 0.05 and the ADF statistic is negative enough to reject the null hypothesis, we can considered the first-order differenced series is stationary.

- However, the plot series shows that the first-order differenced series is still not exhibiting a constant mean and constant variance.

- As such, we might have to apply second differencing on first-order differenced series.

Initial Modelling

Second-order Differencing



```
ADF Statistics: -7.1246
p-value: 0.0000
        1%: -3.633
        5%: -2.949
        10%: -2.613
```

- Since the p-value is below 0.05 and the ADF statistic is negative enough to reject the null hypothesis, we can considered the second-order differenced series is stationary.
- The plot series shows that the second-order differenced series is exhibiting around the constant mean and constant variance.

Initial Modelling

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

ACF

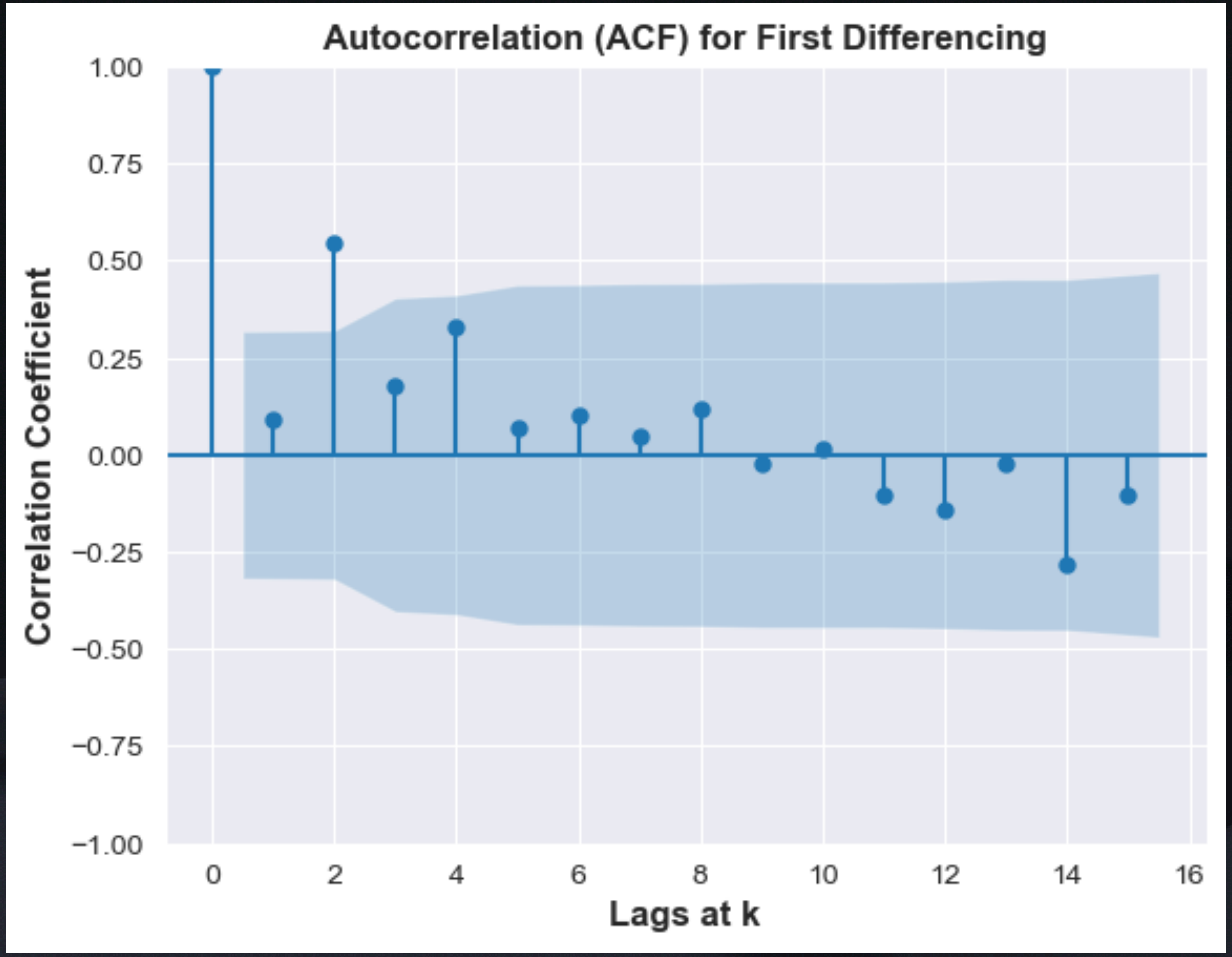
- **Purpose:** It measures the correlation between a time series and its lagged values.
- **Use:** It helps to identify the order of the moving average (MA) component in ARIMA.

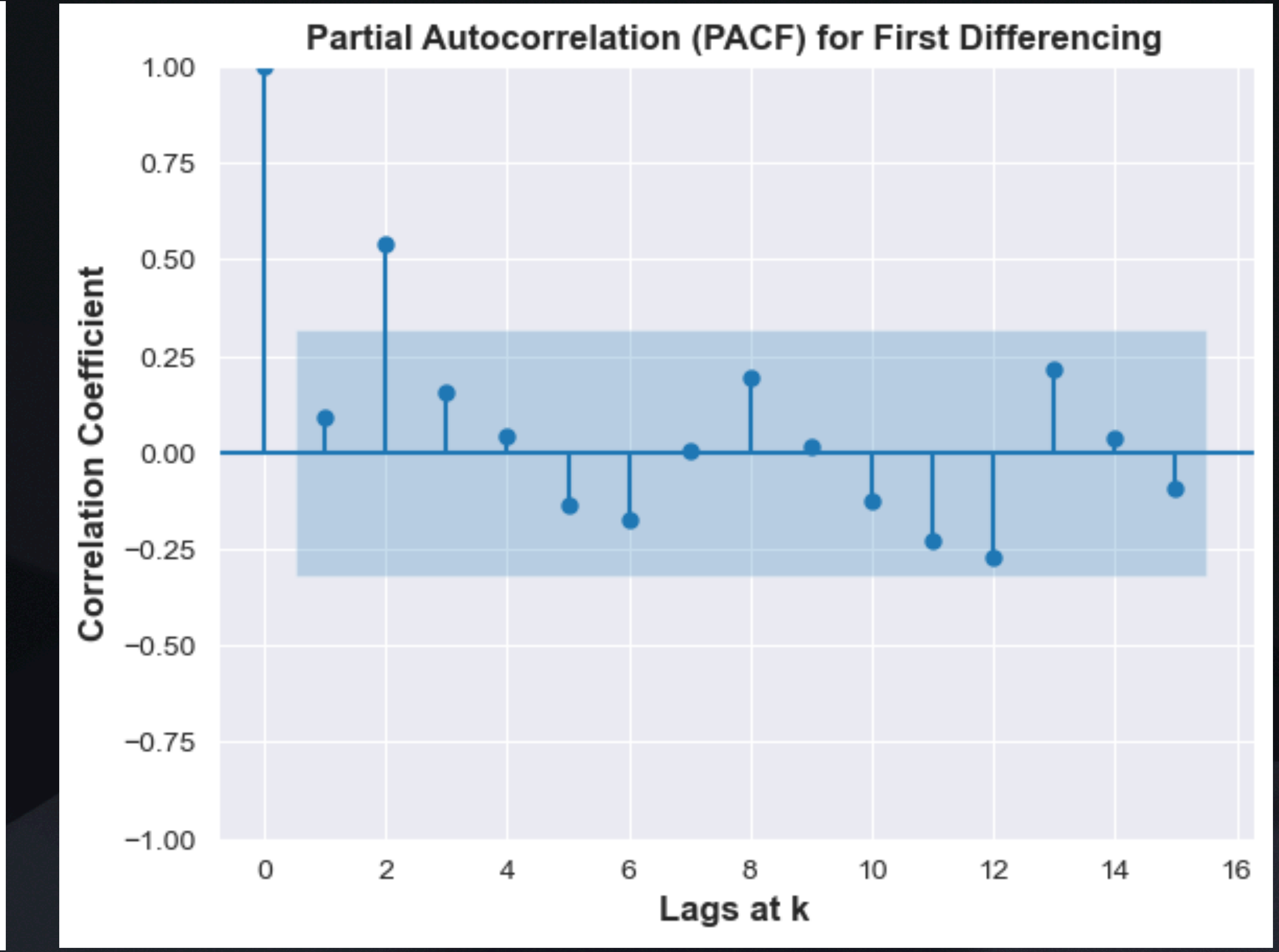
PACF

- **Purpose:** It measures the correlation between a time series and its lagged values after removing the effects of intermediate lags.
- **Use:** It helps to identify the order of the autoregressive (AR) component in ARIMA.

Initial Modelling

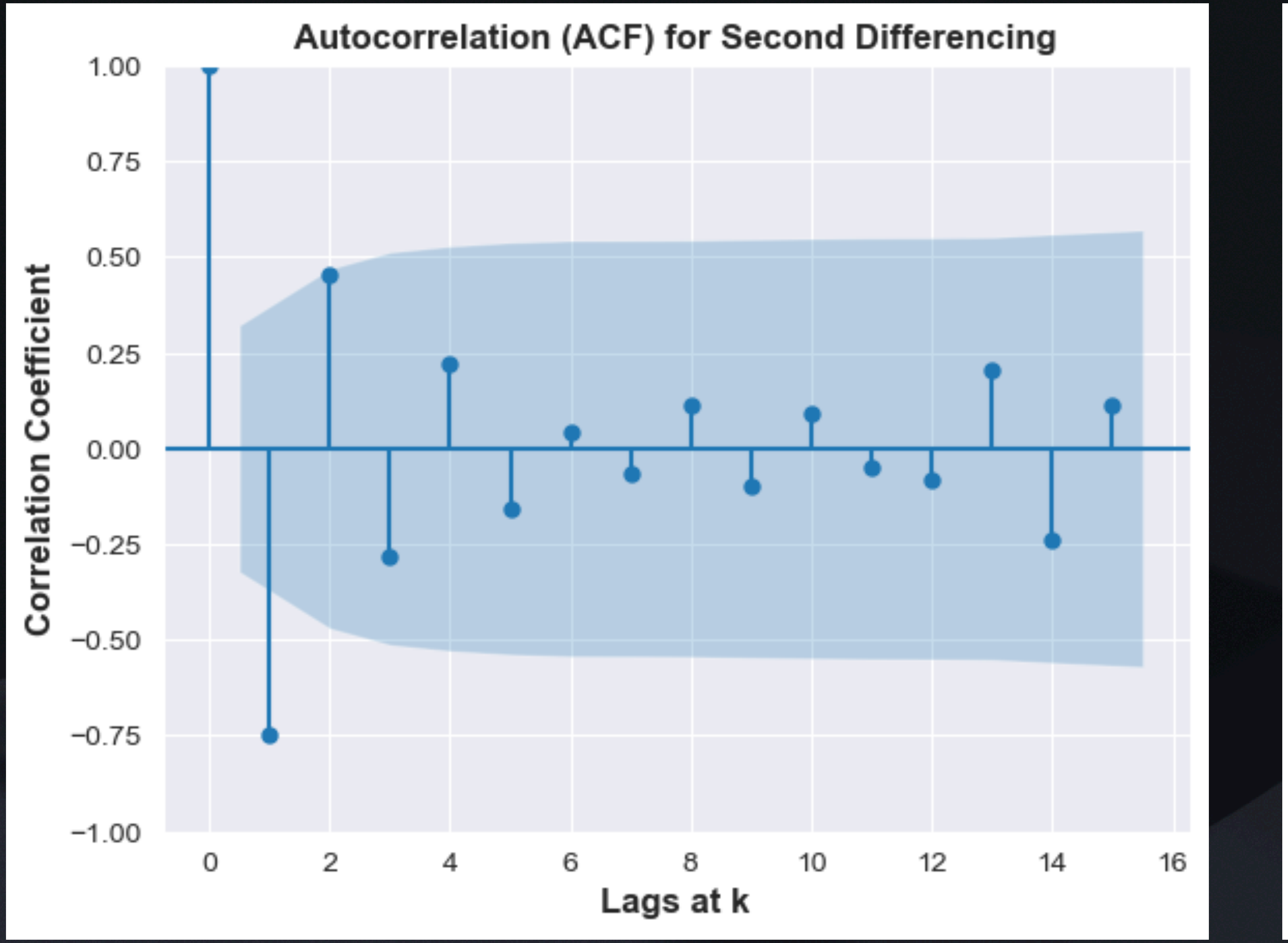
ACF and PACF for First-Order Differencing

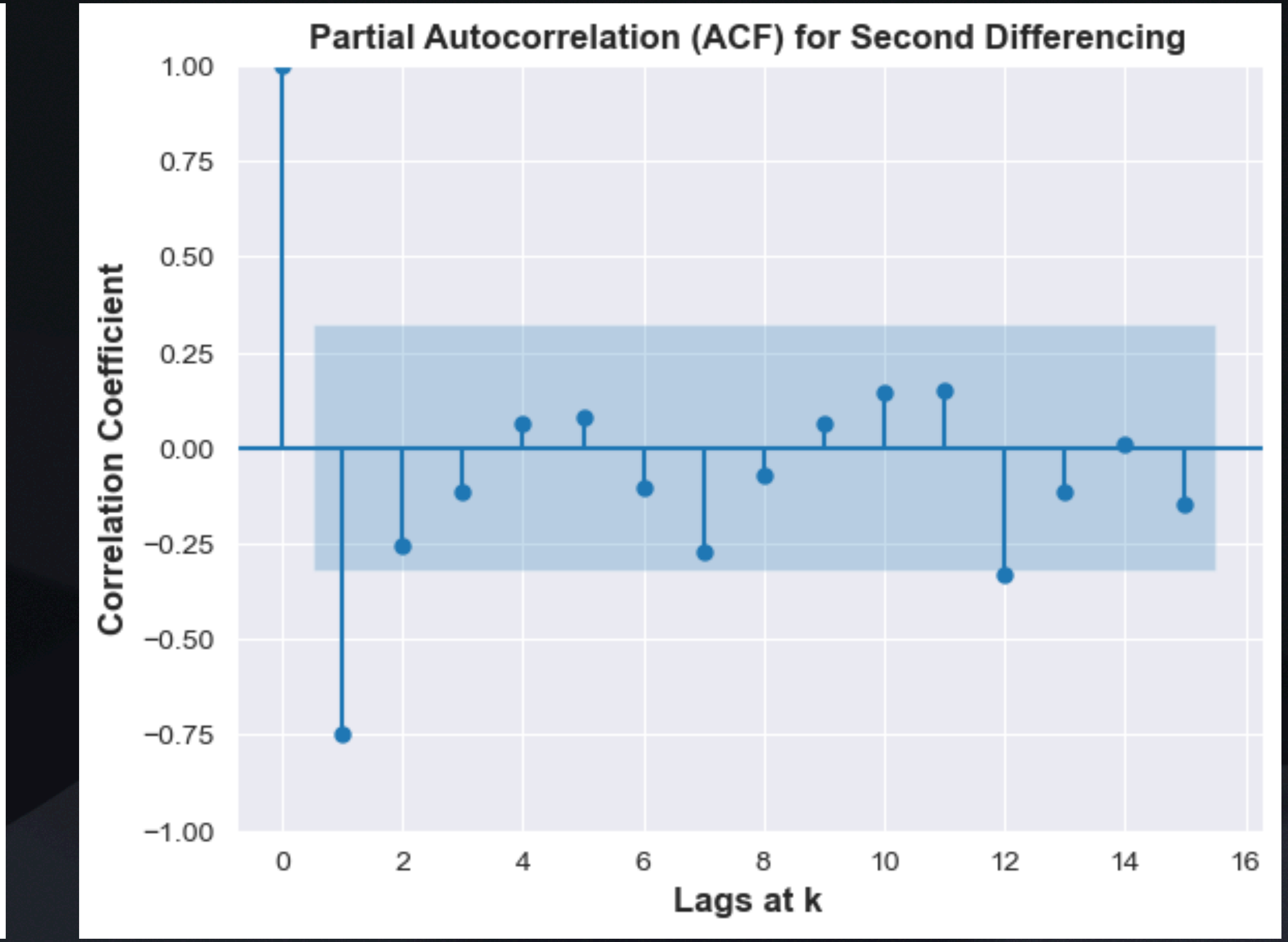




Initial Modelling

ACF and PACF for Second-Order Differencing





Initial Modelling

Possible Parameters for ARIMA

- Based on the **second differencing ACF and PACF plots**:

  - The ACF shows significant negative correlation at lag 1 (or/ and positive correlation at lag 2) and tapers off gradually within the confidence interval. This suggests the presence of moving average (MA) terms.

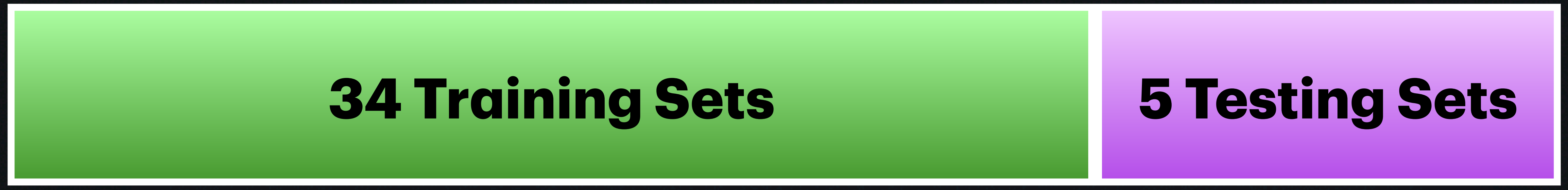
  - The PACF shows a sharp drop-off after lag 1 with other lags mostly within the confidence interval. This suggests the presence of autoregressive (AR) terms.

  - **Possible ARIMA model:** (1, 2, 1) or (2, 2, 1)

Initial Modelling

Manual Configuration ARIMA

- Before fitting into the ARIMA, we split the 39 time series data into 34 training sets and 5 testing sets in a sequential manner.

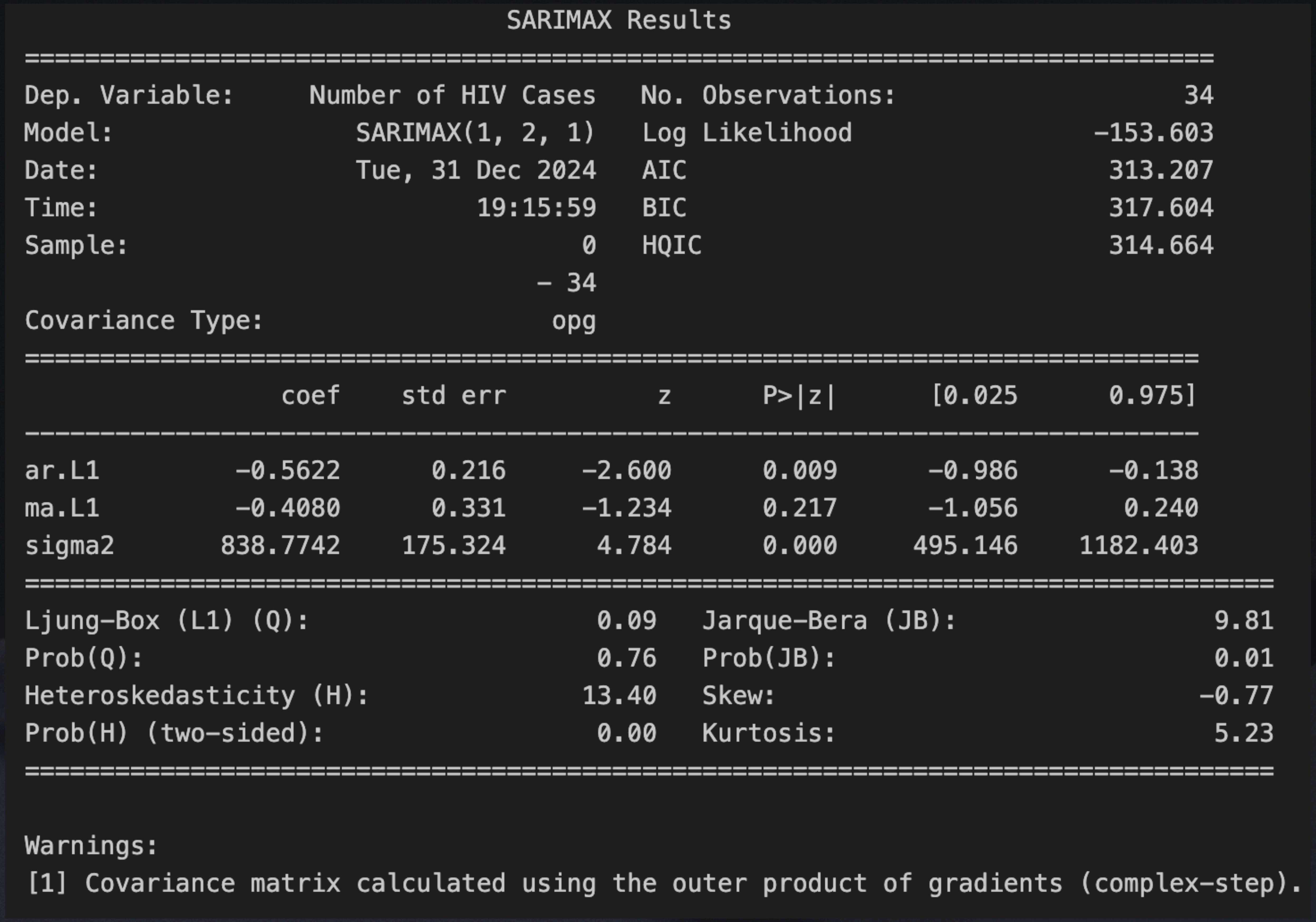


34 Training Sets

5 Testing Sets

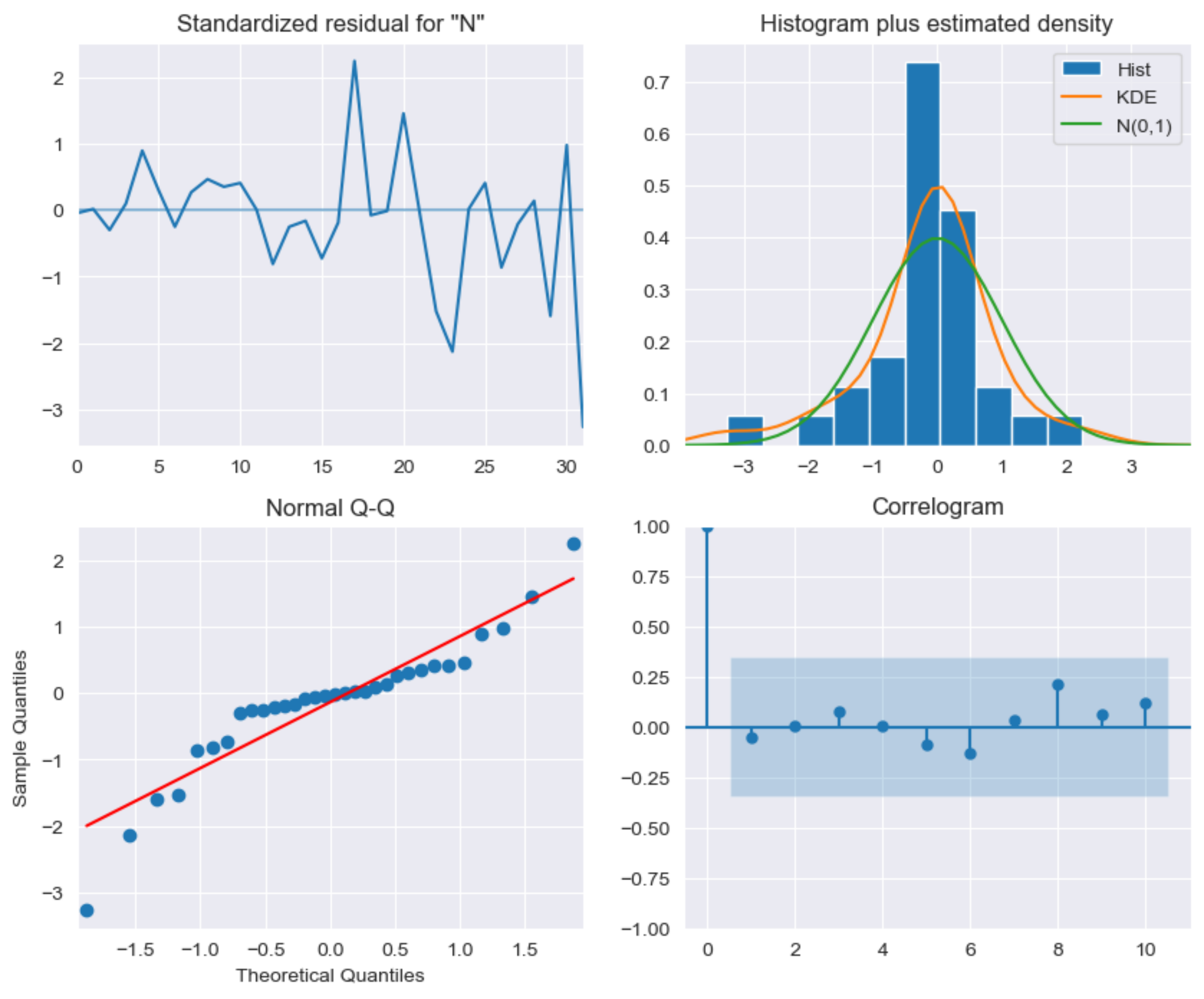
Initial Modelling

Manual Configuration ARIMA (1, 2, 1)



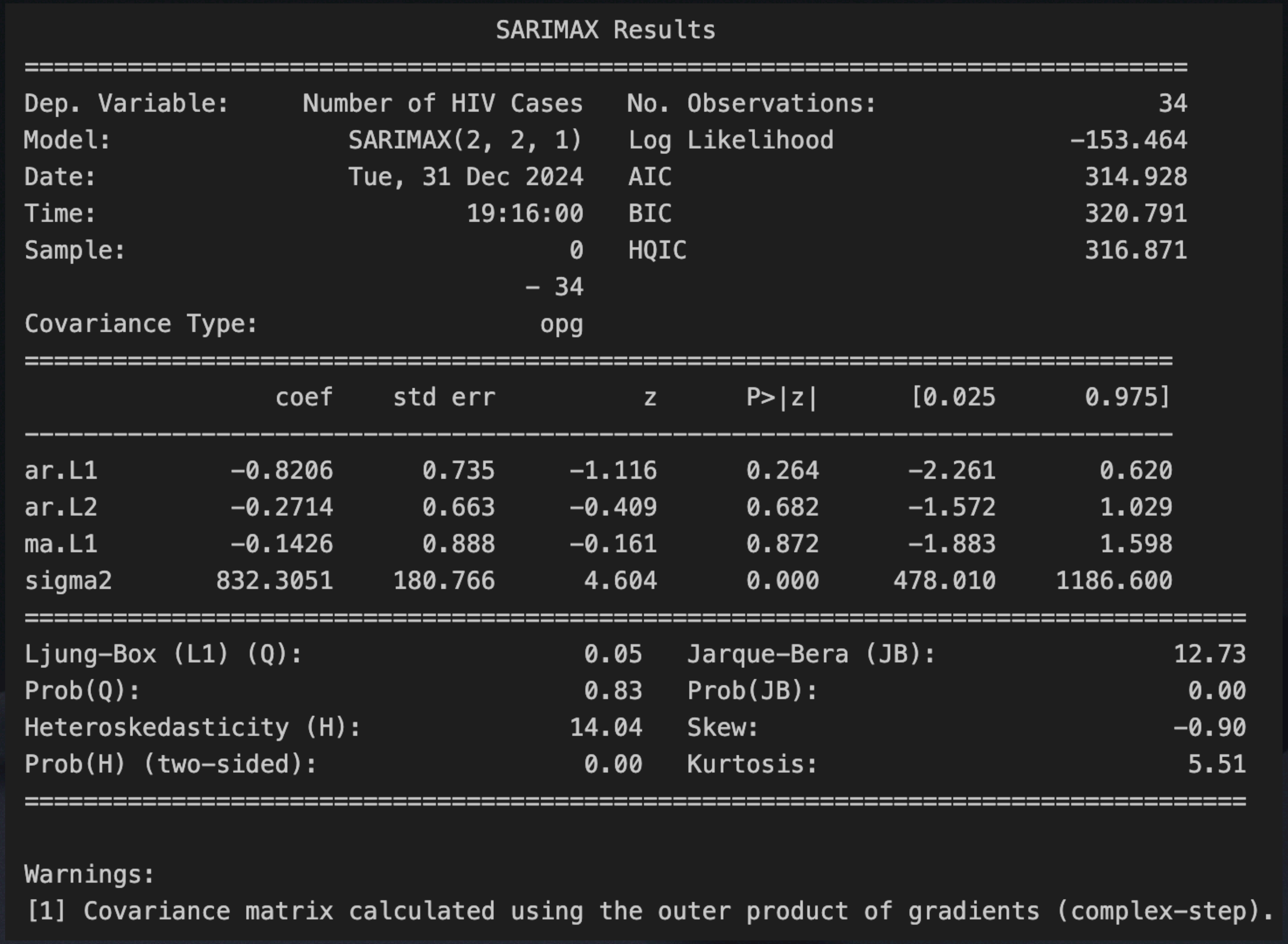
```
                               SARIMAX Results
==============================================================================
Dep. Variable:     Number of HIV Cases   No. Observations:                   34
Model:                SARIMAX(1, 2, 1)   Log Likelihood                -153.603
Date:                 Tue, 31 Dec 2024   AIC                            313.207
Time:                         19:15:59   BIC                            317.604
Sample:                              0   HQIC                           314.664
                                  - 34
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5622      0.216     -2.600      0.009      -0.986      -0.138
ma.L1         -0.4080      0.331     -1.234      0.217      -1.056       0.240
sigma2       838.7742    175.324      4.784      0.000     495.146    1182.403
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 9.81
Prob(Q):                              0.76   Prob(JB):                         0.01
Heteroskedasticity (H):              13.40   Skew:                            -0.77
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.23
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```



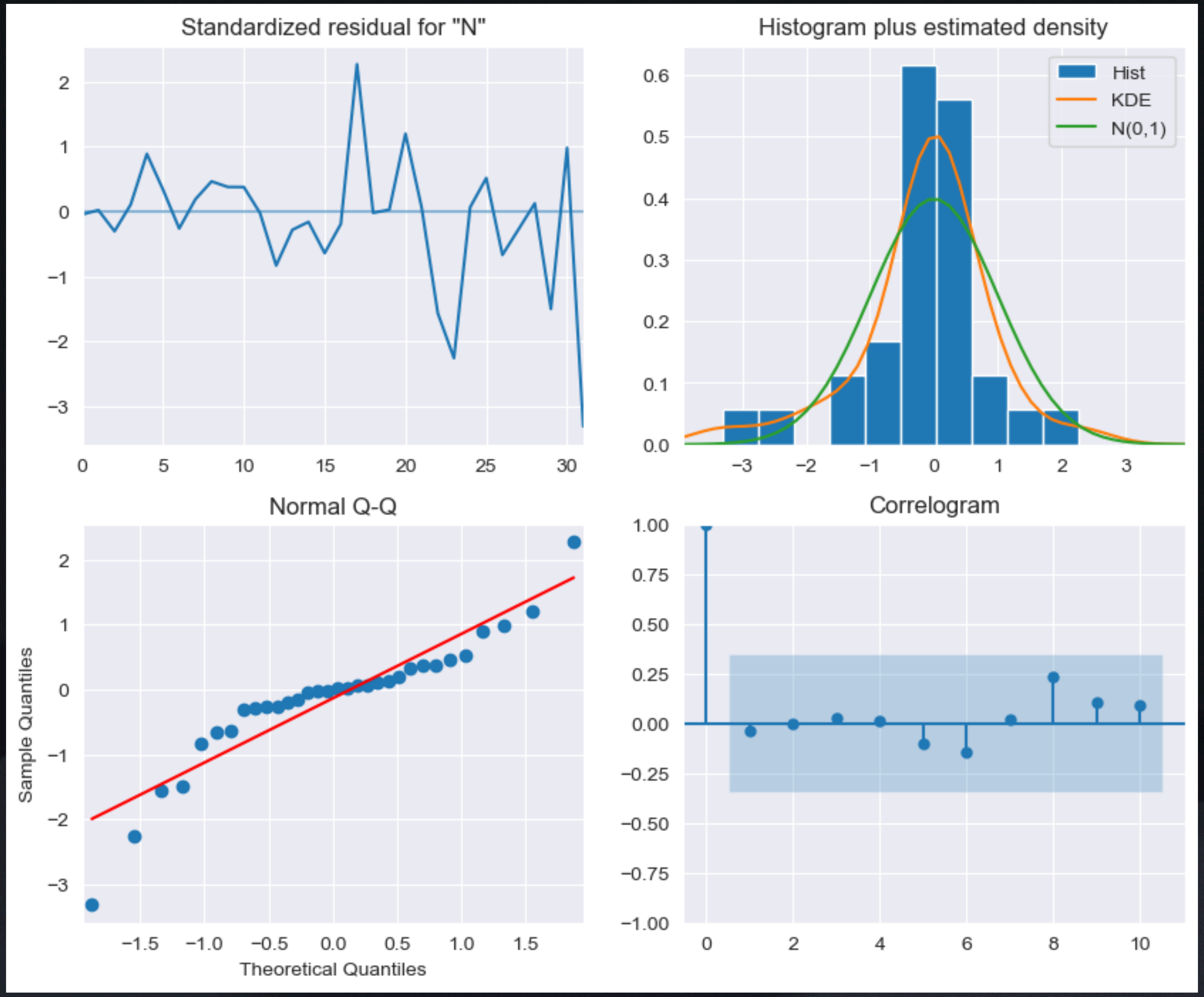
Initial Modelling

Manual Configuration ARIMA (2, 2, 1)



```
                                SARIMAX Results
==============================================================================
Dep. Variable:     Number of HIV Cases   No. Observations:                   34
Model:                SARIMAX(2, 2, 1)   Log Likelihood                -153.464
Date:                 Tue, 31 Dec 2024   AIC                            314.928
Time:                         19:16:00   BIC                            320.791
Sample:                              0   HQIC                           316.871
                                  - 34
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.8206      0.735     -1.116      0.264      -2.261       0.620
ar.L2         -0.2714      0.663     -0.409      0.682      -1.572       1.029
ma.L1         -0.1426      0.888     -0.161      0.872      -1.883       1.598
sigma2       832.3051    180.766      4.604      0.000     478.010    1186.600
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                12.73
Prob(Q):                              0.83   Prob(JB):                         0.00
Heteroskedasticity (H):              14.04   Skew:                            -0.90
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.51
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```



Optimised Modelling

Grid-Search Hyperparameters

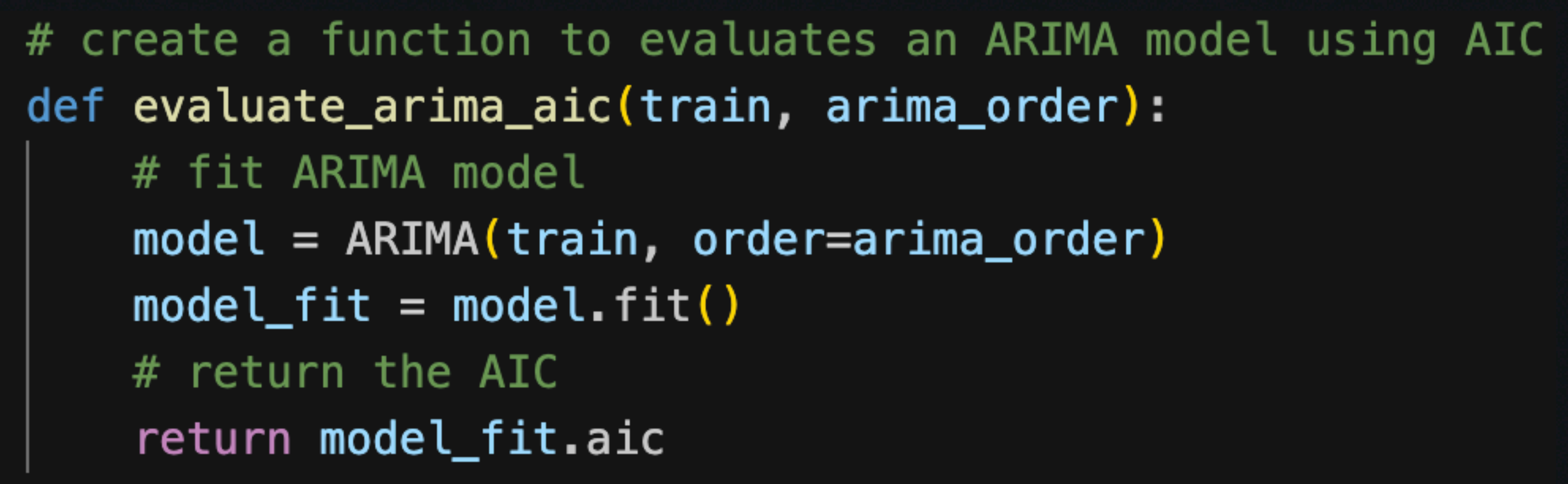
- **Primary Purpose:** A systematic way to explore and find the best combination of ARIMA parameters: p, d, q based on **minimisation of Akaike Information Criterion (AIC).**

- **How it works?**

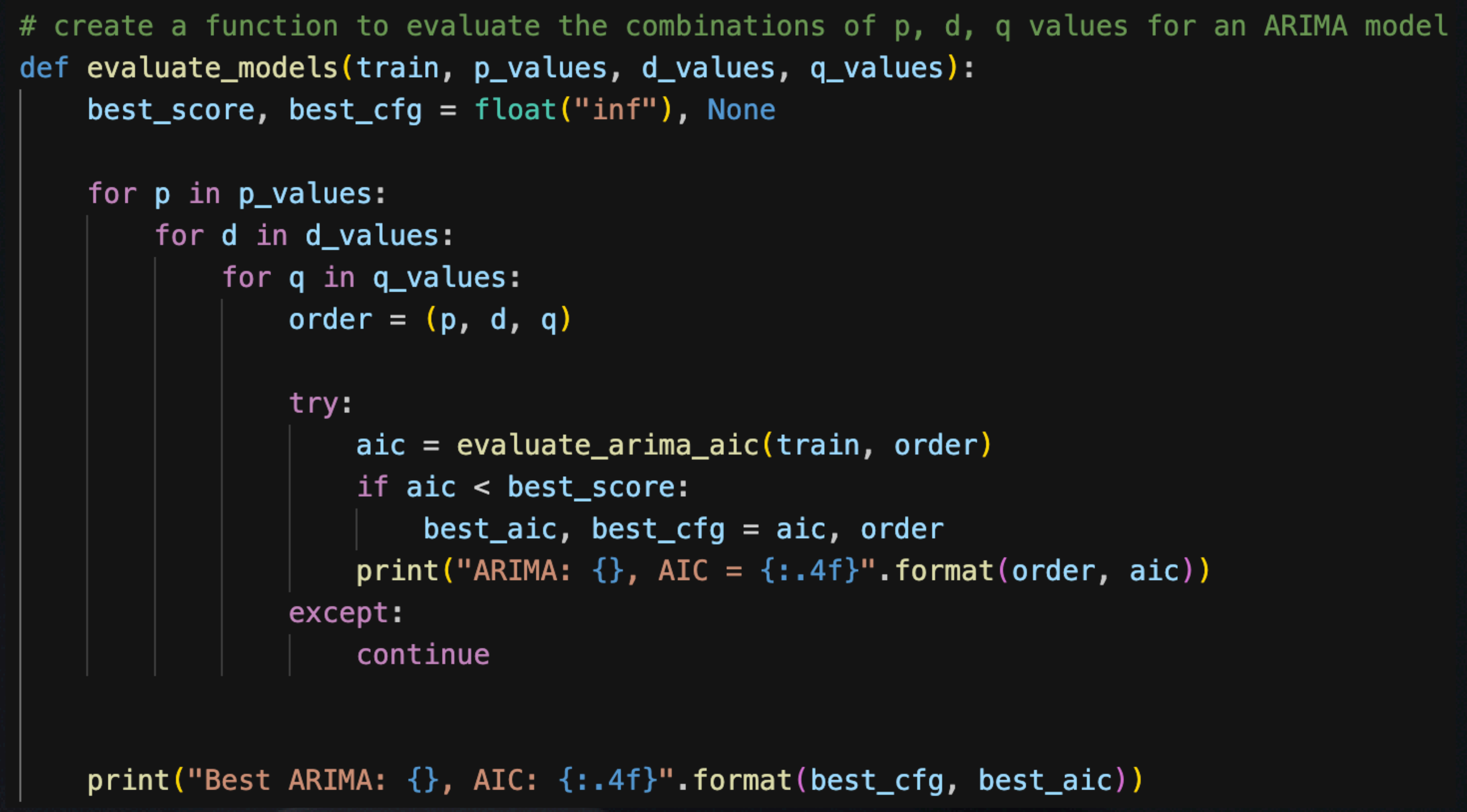
1. Loops through all combinations of p, d, q values.

2. For each combination, fits an ARIMA model using **evaluate_arima_aic** function and then return and records the AIC score.

3. Tracks the combination with the lowest AIC score and display the best ARIMA configuration.



```
# create a function to evaluates an ARIMA model using AIC
def evaluate_arima_aic(train, arima_order):
    # fit ARIMA model
    model = ARIMA(train, order=arima_order)
    model_fit = model.fit()
    # return the AIC
    return model_fit.aic
```

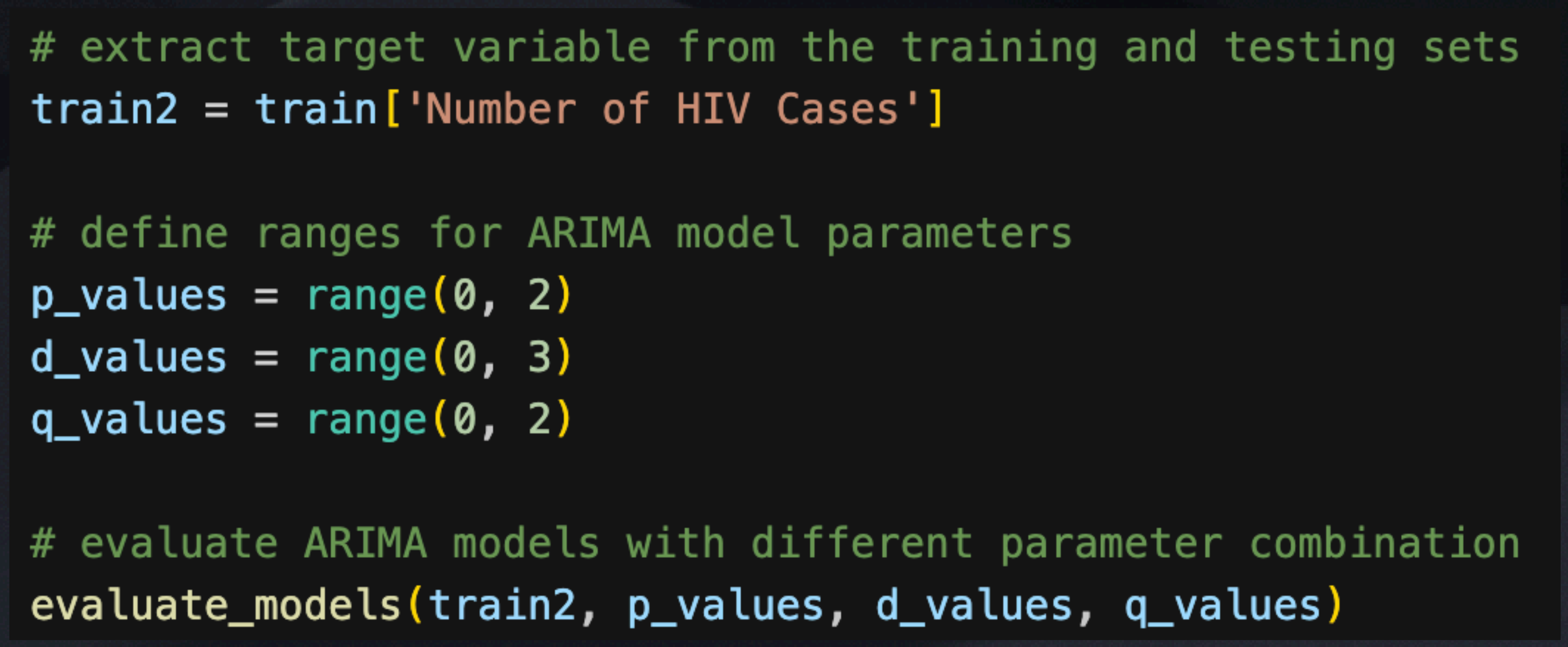


```
# create a function to evaluate the combinations of p, d, q values for an ARIMA model
def evaluate_models(train, p_values, d_values, q_values):
    best_score, best_cfg = float("inf"), None

    for p in p_values:
        for d in d_values:
            for q in q_values:
                order = (p, d, q)

                try:
                    aic = evaluate_arima_aic(train, order)
                    if aic < best_score:
                        best_aic, best_cfg = aic, order
                    print("ARIMA: {}, AIC = {:.4f}".format(order, aic))
                except:
                    continue


    print("Best ARIMA: {}, AIC: {:.4f}".format(best_cfg, best_aic))
```



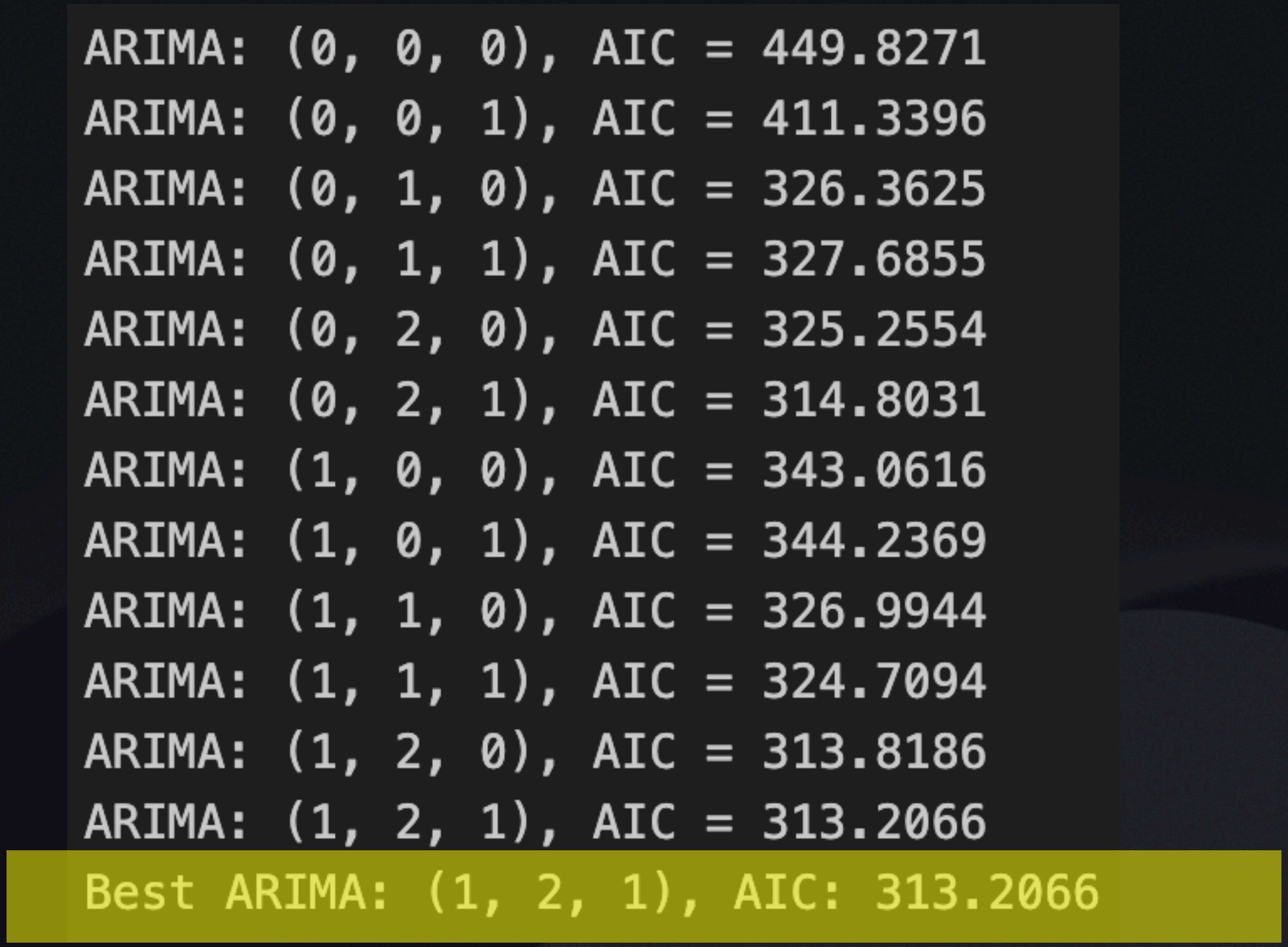
```
# extract target variable from the training and testing sets
train2 = train['Number of HIV Cases']

# define ranges for ARIMA model parameters
p_values = range(0, 2)
d_values = range(0, 3)
q_values = range(0, 2)

# evaluate ARIMA models with different parameter combination
evaluate_models(train2, p_values, d_values, q_values)
```

Optimised Modelling

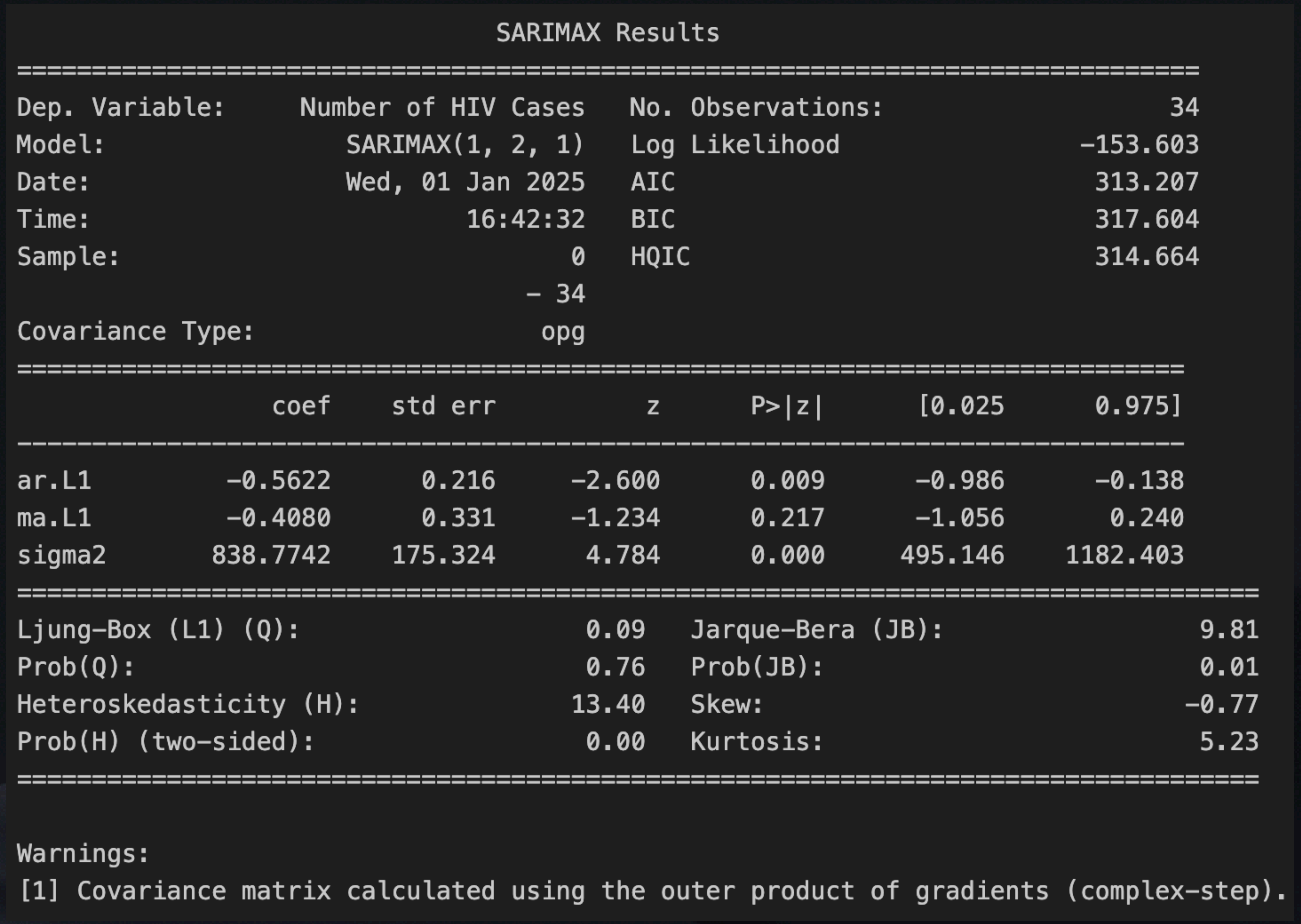
Grid-Search Hyperparameters



```
ARIMA: (0, 0, 0), AIC = 449.8271
ARIMA: (0, 0, 1), AIC = 411.3396
ARIMA: (0, 1, 0), AIC = 326.3625
ARIMA: (0, 1, 1), AIC = 327.6855
ARIMA: (0, 2, 0), AIC = 325.2554
ARIMA: (0, 2, 1), AIC = 314.8031
ARIMA: (1, 0, 0), AIC = 343.0616
ARIMA: (1, 0, 1), AIC = 344.2369
ARIMA: (1, 1, 0), AIC = 326.9944
ARIMA: (1, 1, 1), AIC = 324.7094
ARIMA: (1, 2, 0), AIC = 313.8186
ARIMA: (1, 2, 1), AIC = 313.2066
Best ARIMA: (1, 2, 1), AIC: 313.2066
```

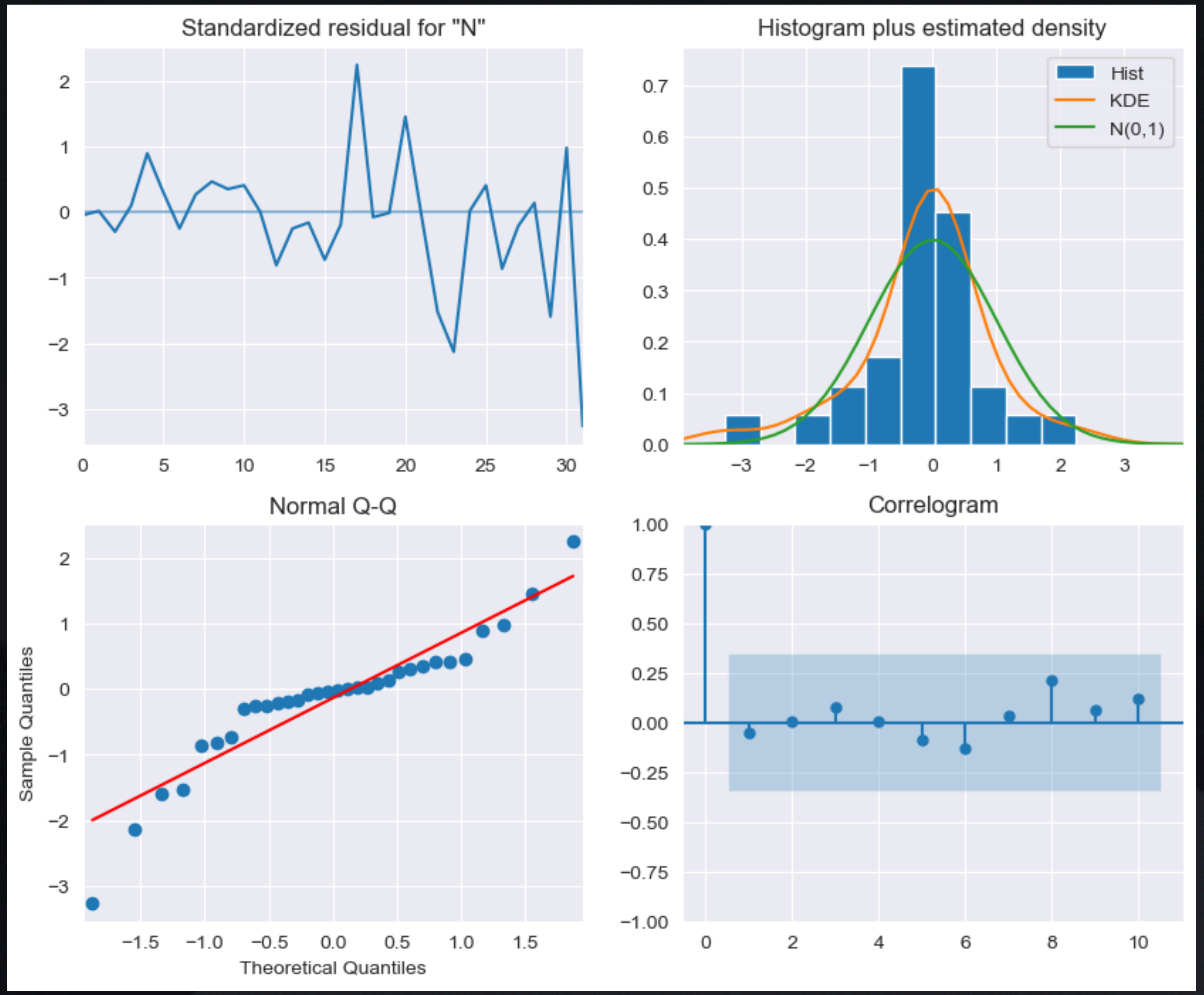
Optimised Modelling

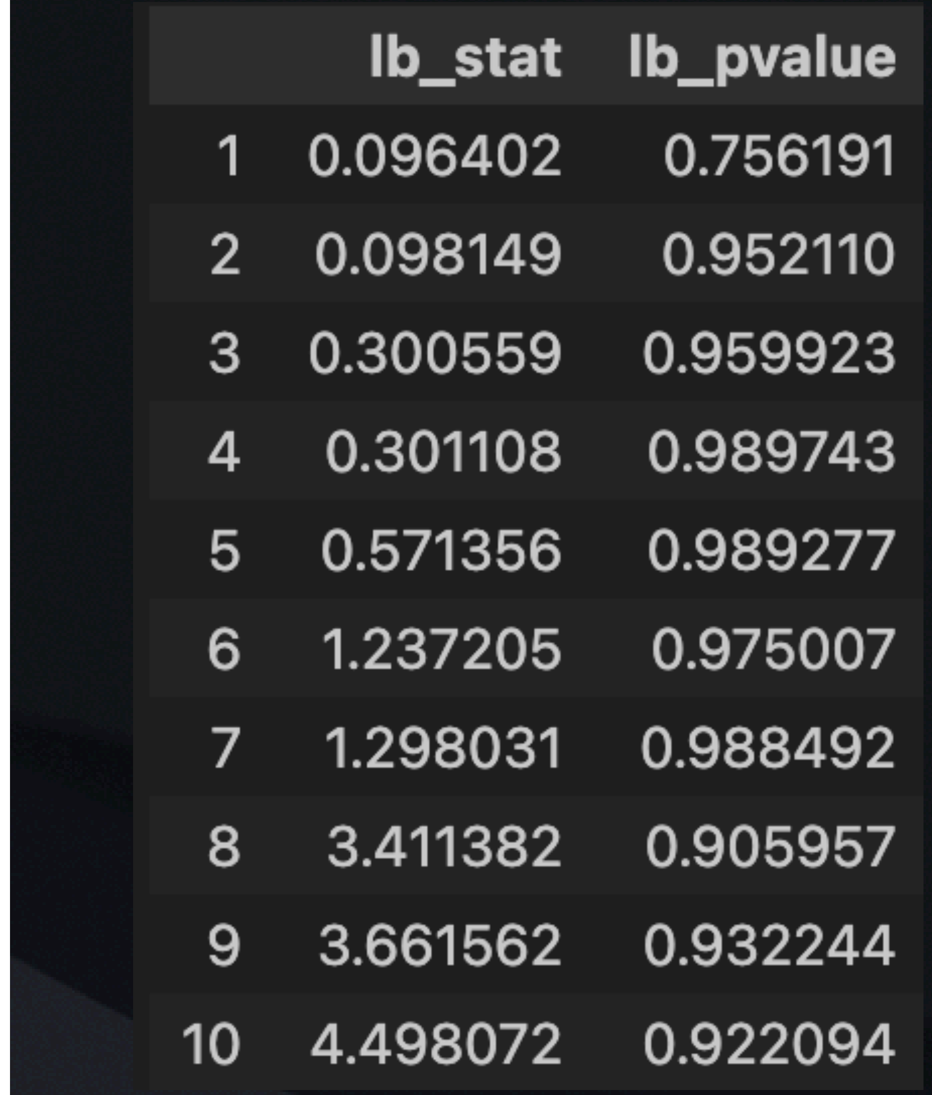
ARIMA (1, 2, 1) - Model Fitting & Evaluation



```
                                   SARIMAX Results
==============================================================================
Dep. Variable:      Number of HIV Cases   No. Observations:                   34
Model:                  SARIMAX(1, 2, 1)   Log Likelihood                -153.603
Date:                   Wed, 01 Jan 2025   AIC                            313.207
Time:                           16:42:32   BIC                            317.604
Sample:                                0   HQIC                           314.664
                                    - 34
Covariance Type:                     opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5622      0.216     -2.600      0.009      -0.986      -0.138
ma.L1         -0.4080      0.331     -1.234      0.217      -1.056       0.240
sigma2       838.7742    175.324      4.784      0.000     495.146    1182.403
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 9.81
Prob(Q):                              0.76   Prob(JB):                         0.01
Heteroskedasticity (H):              13.40   Skew:                            -0.77
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.23
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

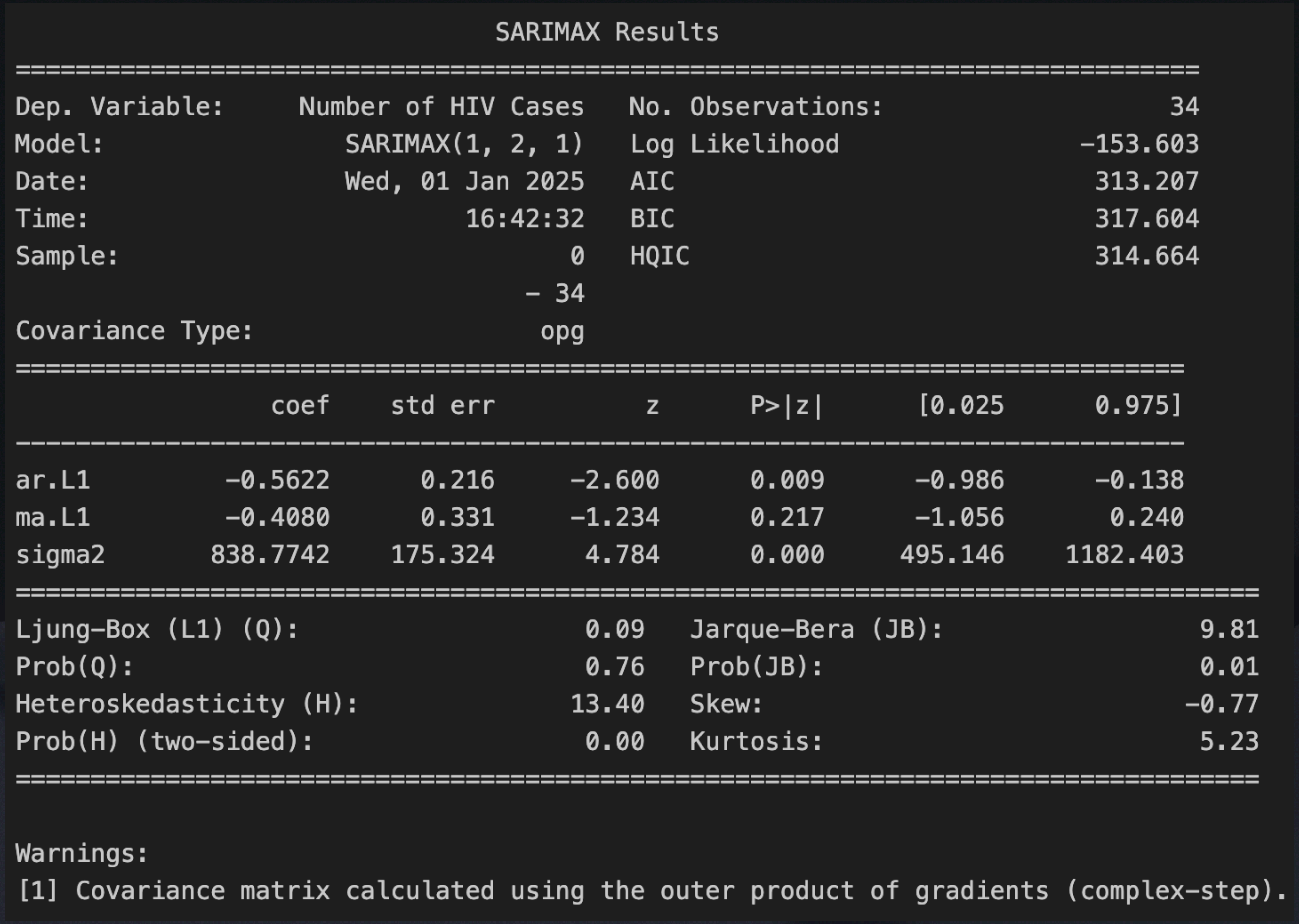




	lb_stat	lb_pvalue
1	0.096402	0.756191
2	0.098149	0.952110
3	0.300559	0.959923
4	0.301108	0.989743
5	0.571356	0.989277
6	1.237205	0.975007
7	1.298031	0.988492
8	3.411382	0.905957
9	3.661562	0.932244
10	4.498072	0.922094

Optimised Modelling

ARIMA (1, 2, 1) - Summary Output Analysis



```
                               SARIMAX Results
==============================================================================
Dep. Variable:      Number of HIV Cases   No. Observations:                   34
Model:                 SARIMAX(1, 2, 1)   Log Likelihood                -153.603
Date:                  Wed, 01 Jan 2025   AIC                            313.207
Time:                          16:42:32   BIC                            317.604
Sample:                               0   HQIC                           314.664
                                   - 34
Covariance Type:                    opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5622      0.216     -2.600      0.009      -0.986      -0.138
ma.L1         -0.4080      0.331     -1.234      0.217      -1.056       0.240
sigma2       838.7742    175.324      4.784      0.000     495.146    1182.403
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 9.81
Prob(Q):                              0.76   Prob(JB):                         0.01
Heteroskedasticity (H):              13.40   Skew:                            -0.77
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.23
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

- **ar.L1:** statistically significant (autoregressive terms contributes to the model)

- **ma.L1:** not statistically significant (moving average may not be critical to the model)

- **sigma2:** statistically significant (measure the variability of the residuals)

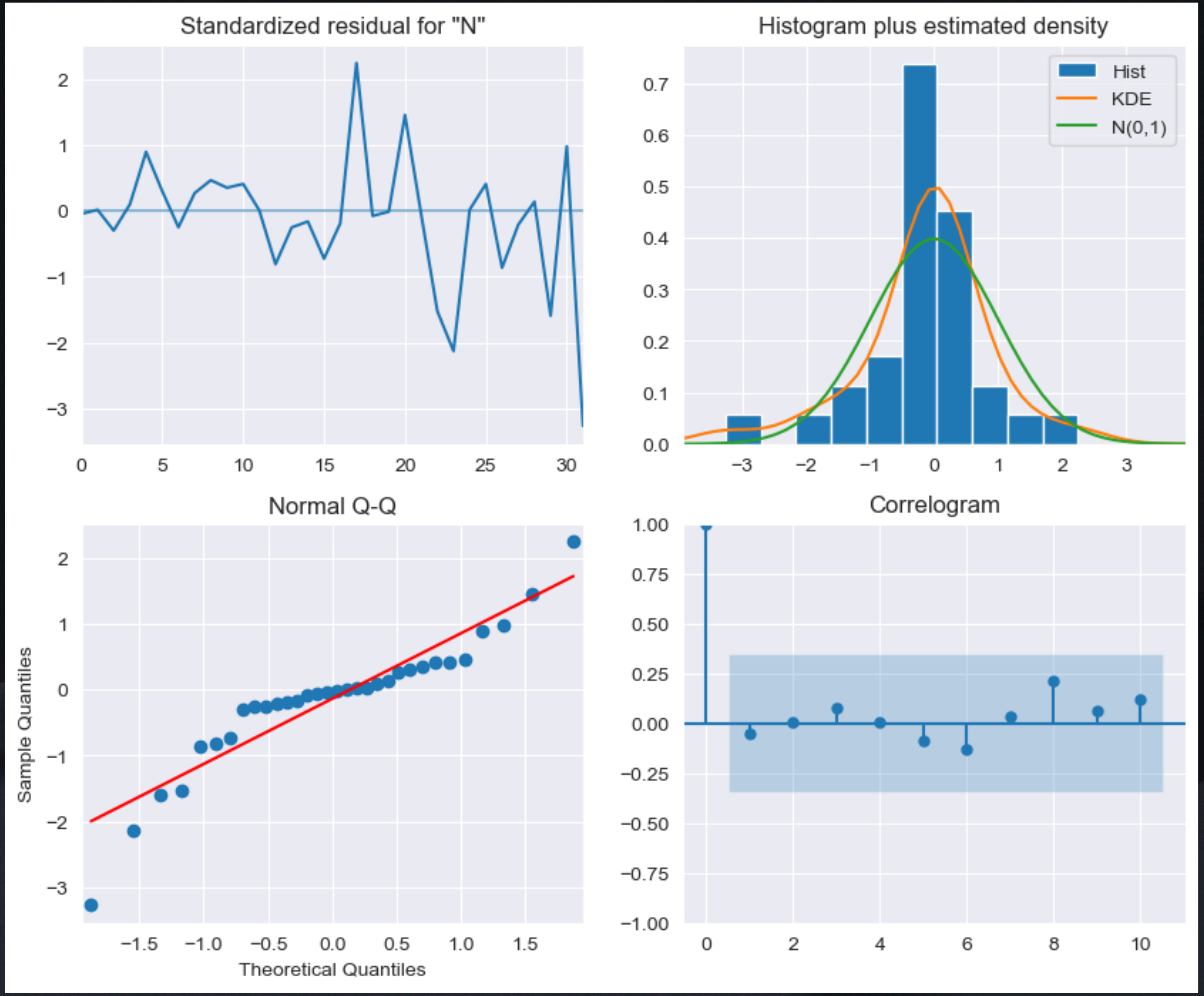
- **AIC:** 313.207

- **BIC:** 317.604

- **HQIC:** 314.664

Optimised Modelling

ARIMA (1, 2, 1) - Diagnostic Plots



- **Standardised Residuals:** The residuals oscillate around zero without obvious patterns, indicating a reasonably well-fitted model.

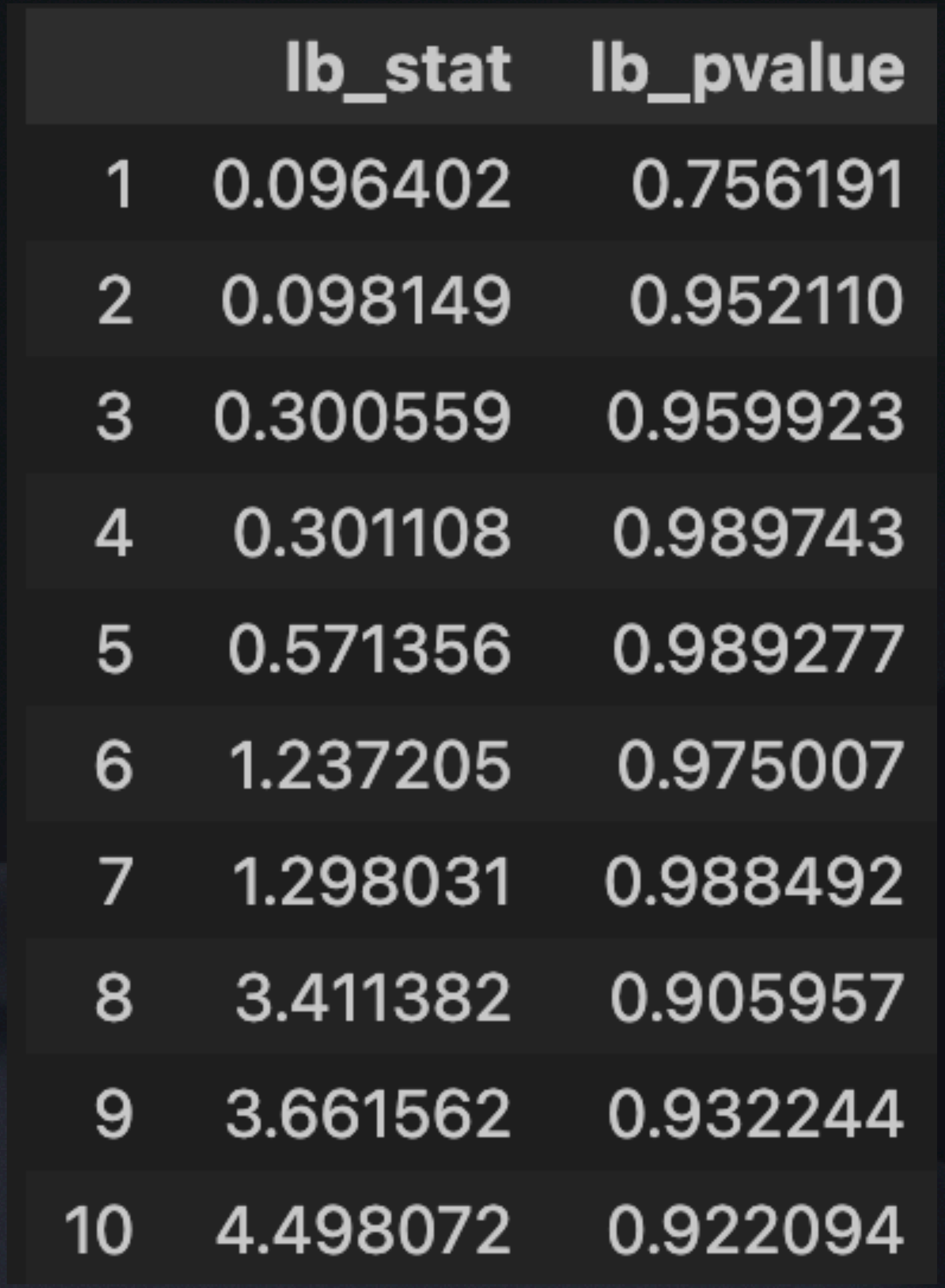
- **Histogram with KDE:** The residuals are approximately normal but show slight skewness and kurtosis.

- **QQ-Plot:** Some deviation from the red line in the tail suggest that residuals are not perfectly normally distributed.

- **Correlogram:** There is no significant correlation in residuals as all lags are lie within the 95% CI.

Optimised Modelling

ARIMA (1, 2, 1) - Ljung-Box Test



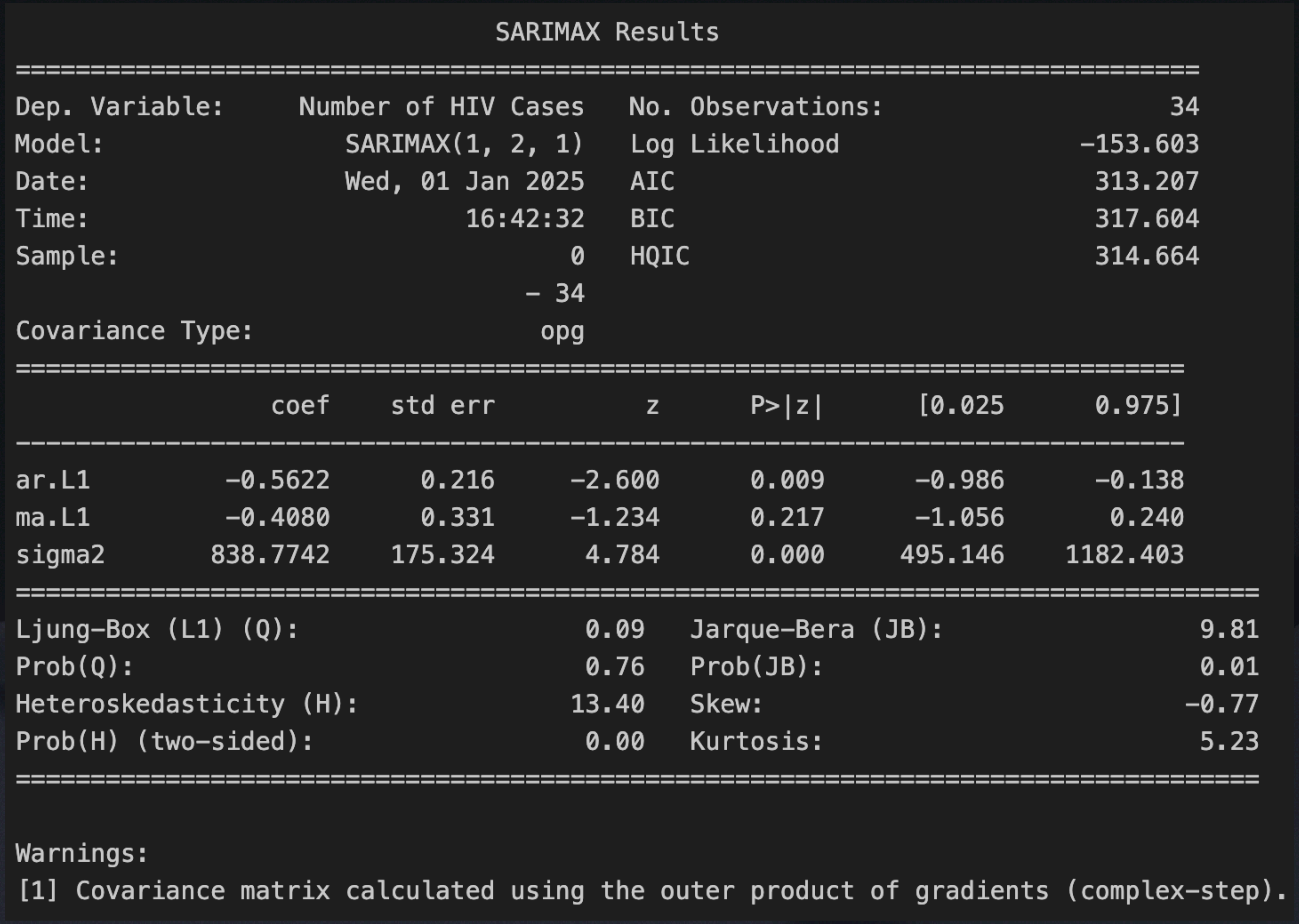
	lb_stat	lb_pvalue
1	0.096402	0.756191
2	0.098149	0.952110
3	0.300559	0.959923
4	0.301108	0.989743
5	0.571356	0.989277
6	1.237205	0.975007
7	1.298031	0.988492
8	3.411382	0.905957
9	3.661562	0.932244
10	4.498072	0.922094

- The Ljung-Box test evaluates the null hypothesis that residuals are independently distributed.

- P-values for all lags are greater than 0.05, which suggests that no evidence of autocorrelation in residuals.

Optimised Modelling

ARIMA (1, 2, 1) - Statistical Tests



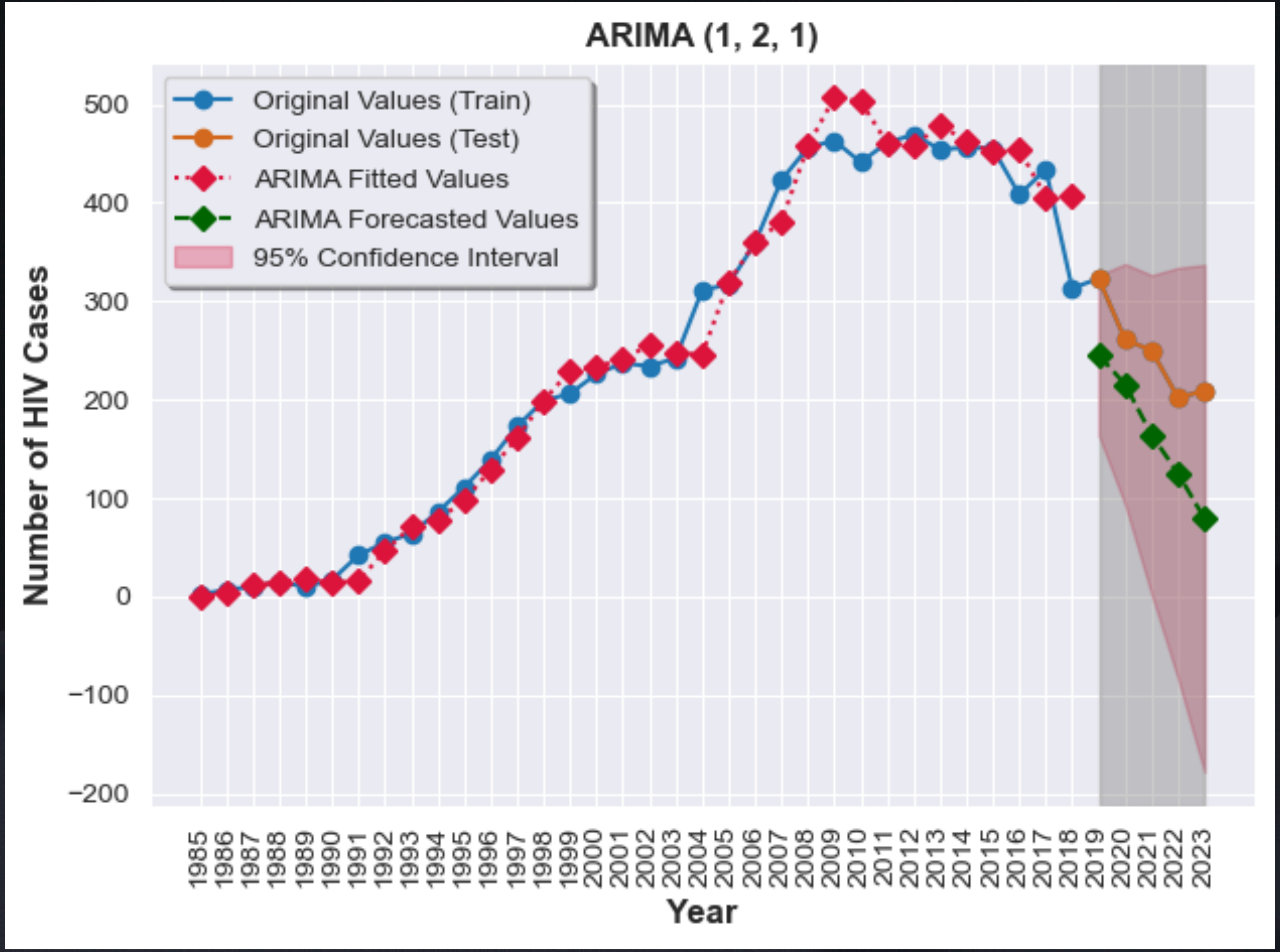
```
                               SARIMAX Results
==============================================================================
Dep. Variable:     Number of HIV Cases   No. Observations:                   34
Model:                SARIMAX(1, 2, 1)   Log Likelihood                -153.603
Date:                 Wed, 01 Jan 2025   AIC                            313.207
Time:                         16:42:32   BIC                            317.604
Sample:                              0   HQIC                           314.664
                                  - 34
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5622      0.216     -2.600      0.009      -0.986      -0.138
ma.L1         -0.4080      0.331     -1.234      0.217      -1.056       0.240
sigma2       838.7742    175.324      4.784      0.000     495.146    1182.403
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 9.81
Prob(Q):                              0.76   Prob(JB):                         0.01
Heteroskedasticity (H):              13.40   Skew:                            -0.77
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.23
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

- **Jarque-Bera Test (JB):** The p-value with 0.01 indicates residuals deviate from normality.
- **Heteroskedasticity (H):** A significant with p-value = 0.00 which suggests heteroskedasticity.
- **Skewness:** -0.77 (deviates from normality)
- **Kurtosis:** 5.23 (deviates from normality)

Optimised Modelling

ARIMA (1, 2, 1) - Forecasting



- The fitted values align closely with the actual values in the training set quite well.

- The forecasted values (green diamond) follow the recent downward trend in the test data (orange points) which indicates the model can reasonably project near-term patterns.

- The forecast 95% confidence interval (shaded red area) widens as the forecast horizon increases which reflecting higher uncertainty over time. This is typical for time series models like ARIMA.

- The confidence interval captures most of the variability in the forecasted area, but it is relatively broad, especially toward the later years, which could limit its usefulness for precise forecasting.

Optimised Modelling

Conclusion

- The optimised ARIMA (1, 2, 1) model effectively captures the overall trend and short-term fluctuations in the number of HIV cases.

- The model diagnostics show reasonable residual behaviour with no significant autocorrelation, adequate goodness-of-fit for both training and testing sets, and reliable short-term forecasts with realistic uncertainty intervals.

- However, the model long-term predictions are less reliable due to widening confidence intervals and potential unrealistic negative values in forecasts.

Optimised Modelling

Future Works

- Develop hybrid models combining ARIMA with ML techniques for better long-term projections.
- Compare ARIMA with alternative models using metrics such as RMSE, MAPE, and AIC.