ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG QUỐC TẾ
VNU-INTERNATIONAL SCHOOL

Part I: Theoretical questions

1) Present the AR(1) model, including its definition; stationarity condition; mean, variance, covariance and autocorrelation functions.
2) Present the 3 estimation methods of the parameters in AR(1) model: method of moments, least squares estimation method and maximum likelihood estimation method.

Part II: Application questions

1) Fit the AR(1) and ARMA(1,1) models for the dataset ***group1_co2_monthly.csv*** and compare the forecast accuracy (RMSE-Root Mean Squared Error and MPAE-Mean Percentage Absolute Error) between these 2 models.
2) By using the AR(1) and ARMA(1,1) models, find the forecasting $\hat{Y}_t(l)$ and 95% prediction limits of Y_{t+l} for the lead time $l = 1$.
3) Check the assumptions of AR(1) model: zero mean and constant variance of errors (by residuals vs. time plot); normality (by Q–Q plot or Shapiro–Wilk test); no autocorrelation (by ACF of residuals or Ljung–Box test).

(The dataset ***group1_co2_monthly.csv*** typically contains monthly atmospheric CO_2 concentrations measured at Mauna Loa Observatory (Hawaii) from 03/1958 to 12/2001).

Questions

Part I: Theoretical questions

1. The AR(1) Model: definition, including its definition; stationarity condition; mean, variance, covariance and autocorrelation functions.

2. Parameter Estimation Methods For The AR(1) Model: MoM, LSE, MLE

Part II: Application Questions

1. 2. 3. *The dataset **group1_co2_monthly.csv** typically contains monthly atmospheric CO_2 concentrations measured at Mauna Loa Observatory (Hawaii) from 03/1958 to 12/2001*

2

The AR(1) Model

- **Definition**

A first-order autoregressive process, abbreviated as AR(1), is defined by the following recursive equation:

$$Y_t = \phi Y_{t-1} + e_t$$

where:

Y_t : the observed time series value at time t

ϕ : the autoregressive coefficient

e_t : white noise, which is a sequence of identically distributed, independent random variables with a mean of zero and variance σ_e^2 .
Crucially, et is assumed to be independent of all previous values of the series $(Y_{t-1}, Y_{t-2}, \ldots)$

The AR(1) Model

- **Stationarity Condition**

  For an AR(1) process to be stationary (meaning its statistical properties do not change over time), it must satisfy the condition:

$$|\phi| < 1$$

  This requirement is often referred to as the stationarity condition.

  If $|\phi| \geq 1$, the process becomes non-stationary:

  - If $|\phi| > 1$, the process is explosive, meaning distant past values have an influence that grows exponentially large, causing the variance and covariance to blow up as time increases.

  - If $|\phi| = 1$, the model becomes a random walk, where the process variance increases linearly with time, and neighboring values remain strongly correlated even as time passes.

The AR(1) Model

- **Mean Function**

Assuming the process has reached statistical equilibrium (stationarity) and no constant intercept is present, the mean function (μ_t) s constant and equal to zero:

$$\mu_1 = E(Y_t) = 0$$

If a non-zero mear μ is desired, the model can be rewritten as

$$(Y_t - \mu) = \phi(Y_{t-1} - \mu) + e_t$$

The AR(1) Model

- **Variance Function**

The process variance, denoted as γ_0 , represents the spread of the data around the mean. For a stationary AR(1) process, it is derived as:

$$\gamma_0 = Var(Y_t) = \frac{\sigma_e^2}{1 - \phi^2}$$

This equation highlights that for the variance to be finite and positive, the condition $\phi^2 < 1 \ (\text{or } |\phi| < 1)$ must hold.

The AR(1) Model

- **Covariance Function**

The Covariance function (γ_k) measures the linear dependence between values in the series separated by k time units (lags). For an AR(1) process, the covariance at lag k is:

$$\gamma_k = Cov(Y_t, Y_{t-k}) = \phi\gamma_{k-1} = \frac{\phi^k \sigma_e^2}{1-\phi^2}$$

This indicates that the covariance at any lag can be built recursively from the previous lag.

The AR(1) Model

Autocorrelation Function (ACF)

The autocorrelation function ρ_k is a unitless measure of dependence, calculated by dividing the autocovariance by the variance $(\rho_k = \frac{\gamma_k}{\gamma_0})$. For an AR(1) model, the ACF is simply:

$$\rho_k = \phi^k \qquad \text{for } k = 1, 2, 3, \ldots$$

The behavior of the ACF depends on the sign of ϕ:

- If $0 < \phi < 1$: All correlations are positive, and the ACF decays exponentially toward zero, resulting in a "smooth" looking series

- If $-1 < \phi < 0$: The signs of successive autocorrelations alternate between positive and negative while their magnitudes decrease exponentially, resulting in a "jagged" series.

Parameter Estimation Methods For The AR(1) Model

- **Method of Moments (MoM)**

The method of moments is typically the simplest approach to obtain parameter estimates by equating sample moments to their theoretical counterparts.

- **Autoregressive Coefficient (ϕ):** For an AR(1) process, the theoretical relationship is $\rho_k = \phi$. Therefore, the method of moments estimate for is simply the lag 1 sample autocorrelation, denoted as: $\widehat{\phi} = r_1$
- **Noise Variance (σ_e^2):** The relationship for the noise variance in an AR(1) process is $\sigma_e^2 = (1 - \rho_1^2)\gamma_0$. By replacing theoretical values with the sample autocorrelation (r_1) and sample variance (s^2), we get: $\sigma_e^2 = (1 - r_1^2)\, s^2$

While these estimates are generally reasonable and easy to compute for autoregressive models, they are considered very inefficient for models containing moving average (MA) terms.

Parameter Estimation Methods For The AR(1) Model

Least Squares Estimation (LSE)

Least squares estimation for an AR(1) model focuses on minimizing the conditional sum-of-squares function, denoted as $S_c(\phi, \mu)$

- **Objective Function:** We minimize the sum of squared differences between observed values and those predicted by the model, starting from the second observation:

$$S_c(\Phi, \mu) = \sum_{t=2}^{n} [(Y_t - \mu) - \Phi(Y_{t-1} - \mu)]^2$$

- **The Estimates:**
  - For the mean (μ), the estimate $\hat{\mu}$ is approximately the sample mean ($\bar{Y}$) for large sample sizes.
  - For the coefficient (ϕ), the resulting estimate is nearly identical to the sample autocorrelation r_1 , differing only by a negligible term in the denominator that disappears as *n* increases.

Parameter Estimation Methods For The AR(1) Model

- **Maximum Likelihood Estimation (MLE)**

  Maximum likelihood estimation uses all information in the data by maximizing the likelihood function, which represents the joint probability density of the observed series.

  - **Unconditional Sum of Squares:** Unlike LSE, MLE typically involves the unconditional sum-of-squares function, $S_c(\phi, \mu)$:

$$S_c(\phi, \mu) = \sum_{t=2}^{n} [(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2 + (1 - \phi^2)(Y_1 - \mu)^2$$

  - **Key Difference:** The primary distinction from LSE is the addition of the term $(1 - \phi^2)(Y_1 - \mu)^2$ which accounts for the probability distribution of the very first observation (Y_1).

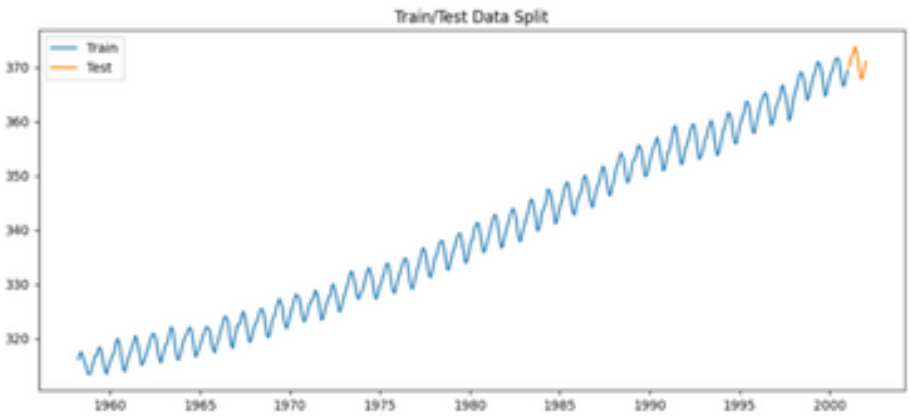
For large samples, MLE and LSE provide nearly identical results. However, MLE is generally more precise, especially when the value of ϕ is near the stationarity boundaries (± 1)

1. **Fit the AR(1) and ARMA(1,1) models for the dataset *group1_co2_monthly.csv* and compare the forecast accuracy (RMSE-Root Mean Squared Error and MPAE-Mean Percentage Absolute Error) between these 2 models.**

The dataset *group1_co2_monthly.csv* contains monthly atmospheric CO_2 concentrations measured at the Mauna Loa Observatory (Hawaii), covering the period from March 1958 to December 2001.

To evaluate the forecasting performance of the model, the dataset is divided into two parts:

- Training set (estimation): all observations except the last 12 months (514 observations)
- Test set (forecasting) : the last 12 monthly observations (12 observations)



- Estimated process mean (intercept): =338.990
- Estimated AR(1) coefficient: =0.9989
- Estimated innovation variance: =1.4122
- The last observed value is =369.44

The plot shows a clear **upward trend** in monthly CO_2 concentrations with a strong and regular seasonal pattern.

1. **Fit the AR(1) and ARMA(1,1) models for the dataset *group1_co2_monthly.csv* and compare the forecast accuracy (RMSE-Root Mean Squared Error and MPAE-Mean Percentage Absolute Error) between these 2 models.**

- **AR(1) model specification and estimation**

The series is modeled using an AR(1) process:

$$Y_t = \mu + \phi(Y_{t-1} - \mu) + e_t$$

→ The model is fitted to the training set using maximum likelihood estimation. The estimated parameters are:

$\hat{\mu}$

$\hat{\phi}$

$\hat{\sigma}^2$

Y_{514}

Specific Fitted Equation: $Y_t = 0.3729 + 0.9978 Y_{t-1}$

1. **Fit the AR(1) and ARMA(1,1) models for the dataset *group1_co2_monthly.csv* and compare the forecast accuracy (RMSE-Root Mean Squared Error and MPAE-Mean Percentage Absolute Error) between these 2 models.**

AR(1) model specification and estimation

To evaluate the forecast performance of the AR(1) model, two commonly used accuracy measures are computed:

- Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \widehat{Y}_t)^2} = 2.4174$$

- Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \widehat{Y}_t}{\widehat{Y}_t}\right| \times 100\% = 0.5424\%$$

	Actual	Predicted	Error
2001-01-31	370.175	369.406279	0.768721
2001-02-28	371.325	369.372594	1.952406
2001-03-31	372.060	369.338947	2.721053
2001-04-30	372.775	369.305338	3.469662
2001-05-31	373.800	369.271765	4.528235
2001-06-30	373.060	369.238229	3.821771
2001-07-31	371.300	369.204730	2.095270
2001-08-31	369.425	369.171269	0.253731
2001-09-30	367.880	369.137844	-1.257844
2001-10-31	368.050	369.104456	-1.054456
2001-11-30	369.375	369.071105	0.303895
2001-12-31	371.020	369.037791	1.982209

Forecast Result

1. **Fit the AR(1) and ARMA(1,1) models for the dataset *group1_co2_monthly.csv* and compare the forecast accuracy (RMSE-Root Mean Squared Error and MPAE-Mean Percentage Absolute Error) between these 2 models.**

- **ARMA(1,1) model specification and estimation**

  The series is modeled using an AR(1) process:

$$Y_t = \mu + \phi(Y_{t-1} - \mu) + \theta e_{t-1} + e_t$$

  → The model is fitted to the training set using maximum likelihood estimation. The estimated parameters are:

  - Estimated process mean (intercept): $\hat{\mu}$ =339.2916
  - Estimated AR(1) coefficient: $\hat{\phi}$=0.9979
  - Estimated MA(1) coefficient : θ = 0.6571
  - Estimated innovation variance: $\hat{\sigma}^2$ =0.7760

  Specific Fitted Equation: $Y_t = 0.7125 + 0.9978Y_{t-1} + 0.6571\, e_{t-1}$

1. **Fit the AR(1) and ARMA(1,1) models for the dataset *group1_co2_monthly.csv* and compare the forecast accuracy (RMSE-Root Mean Squared Error and MPAE-Mean Percentage Absolute Error) between these 2 models.**

- **ARMA(1,1) model specification and estimation**

To evaluate the forecast performance of the AR(1) model, two commonly used accuracy measures are computed:

- Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2} = 2.1908$$

- Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{\hat{Y}_t}\right| \times 100\% = 0.4806\%$$

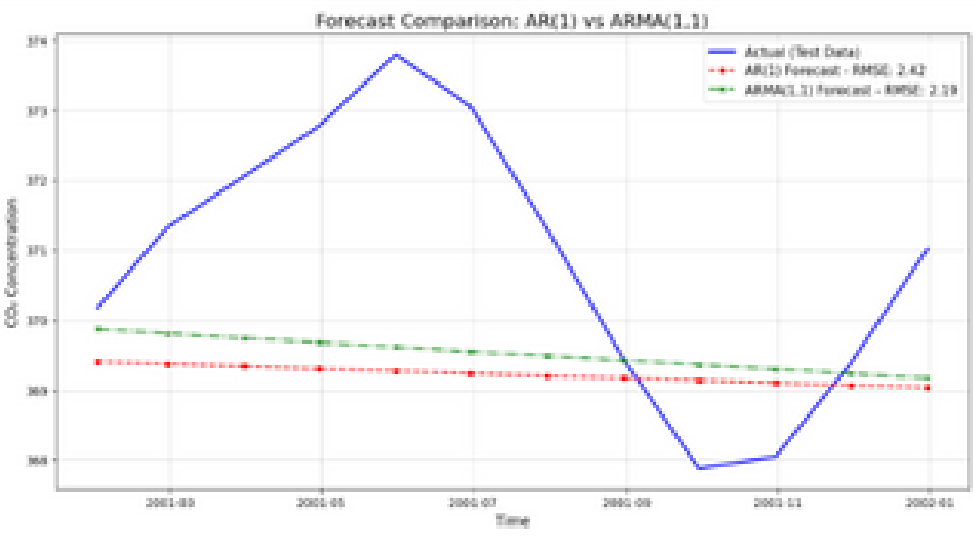
	Actual	Predicted	Error
2001-01-31	370.175	369.872950	0.302050
2001-02-28	371.325	369.808679	1.516321
2001-03-31	372.060	369.744544	2.315456
2001-04-30	372.775	369.680543	3.094457
2001-05-31	373.800	369.616677	4.183323
2001-06-30	373.060	369.552945	3.507055
2001-07-31	371.300	369.489347	1.810653
2001-08-31	369.425	369.425882	-0.000882
2001-09-30	367.880	369.362551	-1.482551
2001-10-31	368.050	369.299354	-1.249354
2001-11-30	369.375	369.236288	0.138712
2001-12-31	371.020	369.173356	1.846644

Forecast Result

1. **Fit the AR(1) and ARMA(1,1) models for the dataset *group1_co2_monthly.csv* and compare the forecast accuracy (RMSE-Root Mean Squared Error and MPAE-Mean Percentage Absolute Error) between these 2 models.**

- **Comparison of AR(1) and ARMA(1,1) Models**

Model	RMSE	MAPE (%)
AR (1)	2.4174	0.5424
ARMA(1,1)	2.1908	0.4806



The ARMA(1,1) model achieves better forecast accuracy than the AR(1) model, as evidenced by lower RMSE and MAPE values.

2. By using the AR(1) and ARMA(1,1) models, find the forecasting $\hat{Y}_t(l)$ and 95% prediction limits of Y_{t+l} for the lead time l=1 .

The AR(1) model

Using the full dataset, the time series is modeled by an AR(1) process:

$$Y_t = \mu + \phi(Y_{t-1} - \mu) + e_t$$

- Estimated process mean (intercept): $\hat{\mu}$ =339.6162
- Estimated AR(1) coefficient: $\hat{\phi}$ =0.9990
- Estimated innovation variance: $\hat{\sigma}^2$ =1.4142
- The last observed value is Y_{526} =369.44

Specific Fitted Equation: $Y_t = 0.3396 + 0.9990\, Y_{t-1}$

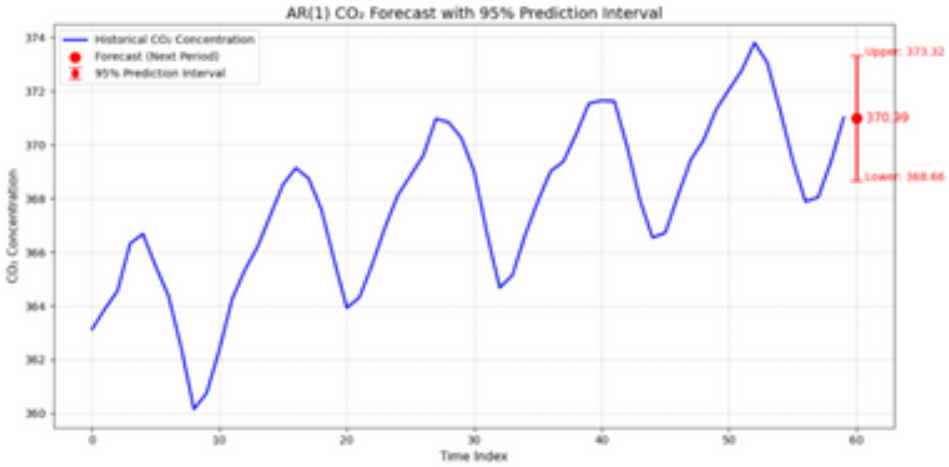
$$\hat{Y}_{t+1} = 370.9881$$

2. **By using the AR(1) and ARMA(1,1) models, find the forecasting $\widehat{Y}_t(l)$ and 95% prediction limits of Y_{t+l} for the lead time l=1 .**

The AR(1) model

Based on the estimation results, the 95% prediction limits for l = 1 are:

- Lower bound (95% CI): 368.6574
- Upper bound (95% CI): 373.3189
- Confidence interval width (±): 2.3308



The point forecast suggests that the CO_2 concentration at time t+1 is expected to be approximately 370.99.
→ The relatively narrow 95% prediction interval indicates high short-term predictability, which is consistent with the near-unit-root behavior of the estimated AR(1) model.

2. **By using the AR(1) and ARMA(1,1) models, find the forecasting $\hat{Y}_t(l)$ and 95% prediction limits of Y_{t+l} for the lead time l=1 .**

- **The ARMA(1,1) model**

Using the full dataset, the time series is modeled by an ARMA(1,1) process:

$$Y_t = \mu + \phi(Y_{t-1} - \mu) + \theta e_{t-1} + e_t$$

- Estimated process mean (intercept): $\hat{\mu}$ =340.0583
- Estimated AR(1) coefficient: $\hat{\phi}$ =0.9981
- Estimated innovation variance: $\hat{\sigma}^2$ = 0.7745
- The Estimated MA(1) coefficient is $\hat{\theta}$=0.6602

Specific Fitted Equation: $Y_t = 0.6461 + 0.9981 Y_{t-1} + 0.6602\, e_{t-1}$

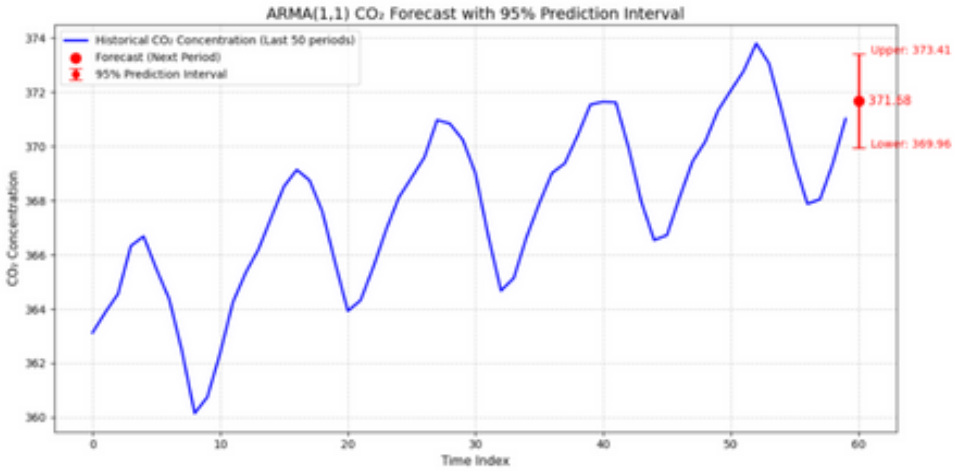
$$\Rightarrow \quad \hat{Y}_{t+1} = 370.9881$$

2. **By using the AR(1) and ARMA(1,1) models, find the forecasting $\widehat{Y}_t(l)$ and 95% prediction limits of Y_{t+l} for the lead time l=1 .**

- **The ARMA(1,1) model**

Based on the estimation results, the 95% prediction limits for l = 1 are:

- Lower bound (95% CI): 369.9577
- Upper bound (95% CI): 373.4073
- Confidence interval width (±): 3.4497



The ARMA(1,1) model produces a point forecast of approximately 371.68 for the next period t+1

→ The prediction interval quantifies the uncertainty associated with the one-step-ahead forecast and reflects the contribution of both autoregressive and moving-average components.

3. **Check the assumptions of AR(1) model: zero mean and constant variance of errors (by residuals vs. time plot); normality (by Q–Q plot or Shapiro–Wilk test); no autocorrelation (by ACF of residuals or Ljung–Box test).**

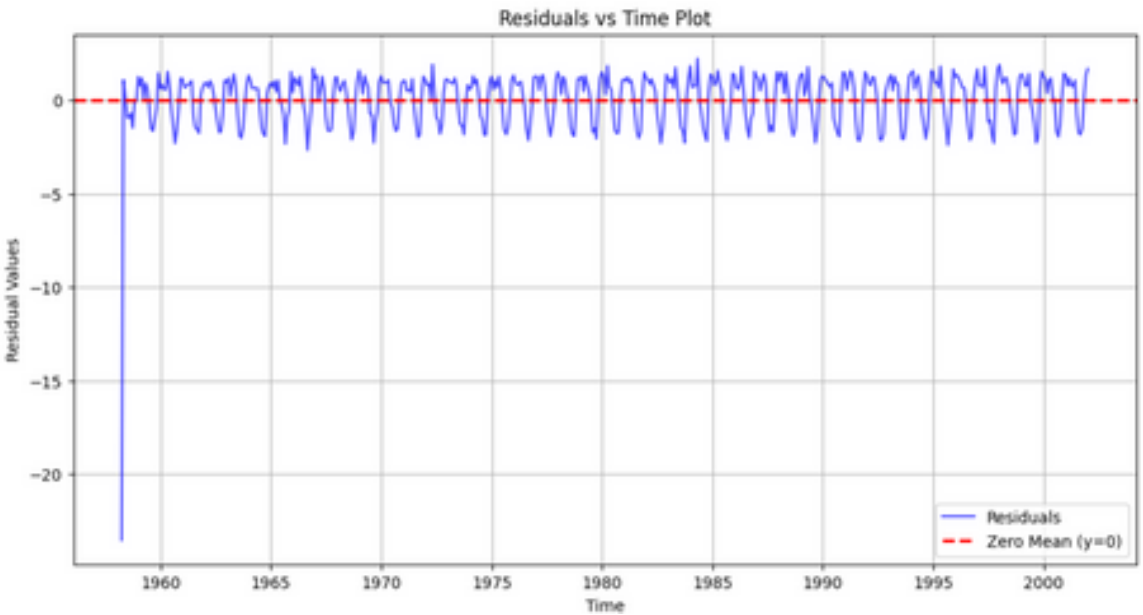
Assumption 1: Zero mean and constant variance of errors

Zero Mean of Errors

Let e_t denote the residuals of the AR(1) model.

- H_0: $E(e_t) = 0$ (the residuals have zero mean)
- H_1: $E(e_t) \neq 0$ (the residuals do not have zero mean)

The mean of the residuals is estimated to be **0.0596** (<0.1), which is very close to 0



In the residuals vs time plot, the residuals fluctuate randomly around the horizontal zero line without exhibiting systematic bias or persistent deviation. This visual pattern indicates that positive and negative residuals balance each other over time.

→ *Therefore,* ***there is no evidence to reject H_0****, and the assumption that the error terms have zero mean is satisfied.*

Zero Mean: PASS

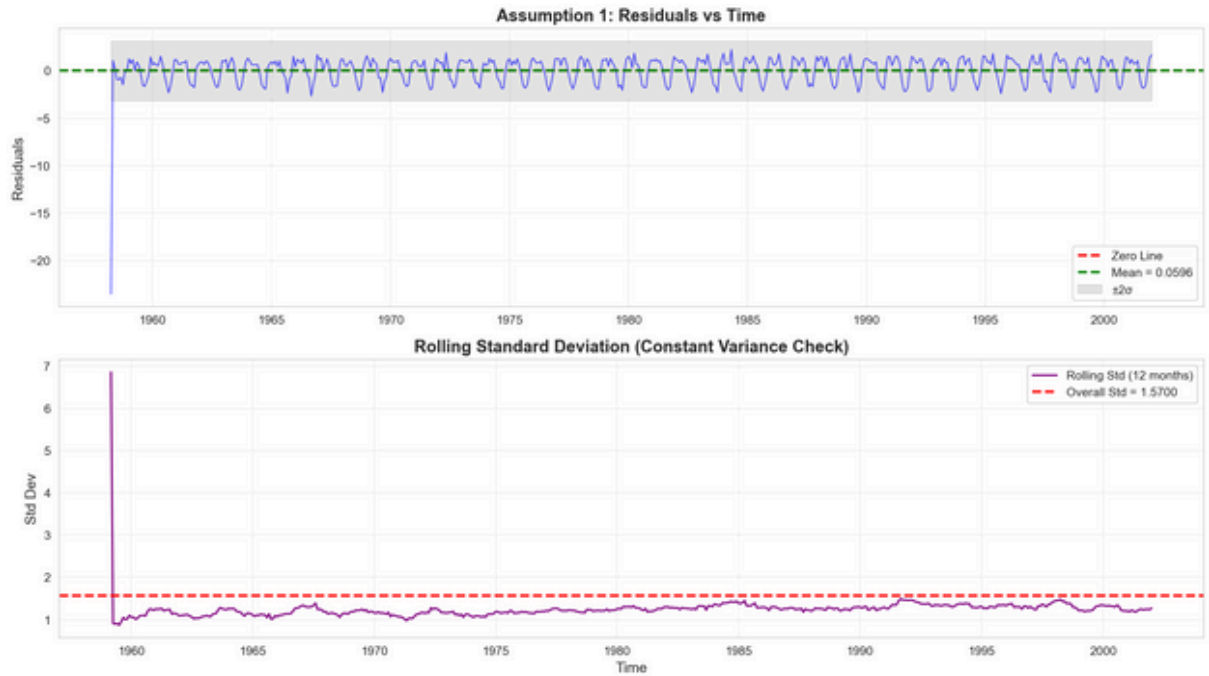
3. **Check the assumptions of AR(1) model: zero mean and constant variance of errors (by residuals vs. time plot); normality (by Q–Q plot or Shapiro–Wilk test); no autocorrelation (by ACF of residuals or Ljung–Box test).**

Assumption 1: Zero mean and constant variance of errors

- **Constant Variance of Errors (Homoscedasticity)**

$$H_0: Var(e_t) = \sigma^2 \qquad H_1: Var(e_t) \text{ is not constant over time}$$

The standard deviation of the residuals is **1.5700**, and the residuals vs. time plot shows a relatively stable spread throughout the entire sample period.



Residuals vs time (top); Rolling standard deviation (bottom)

There is no visible pattern of increasing or decreasing variability, nor any volatility clustering. The dispersion of residuals appears homogeneous across time.

*→ Consequently, **H_0 is not rejected**, indicating that the constant variance assumption holds.*

Based on the residual diagnostics, the AR(1) model satisfies both the zero mean and constant variance assumptions of the error terms. These results support the adequacy of the AR(1) model from the perspective of basic error assumptions.

Constant Variance: PASS

3. **Check the assumptions of AR(1) model: zero mean and constant variance of errors (by residuals vs. time plot); normality (by Q–Q plot or Shapiro–Wilk test); no autocorrelation (by ACF of residuals or Ljung–Box test).**

Assumption 2: Normality of Errors

The Q–Q plot reveals clear deviations of the residuals from the reference line, especially in the lower tail, suggesting a departure from normality and the presence of extreme observations.

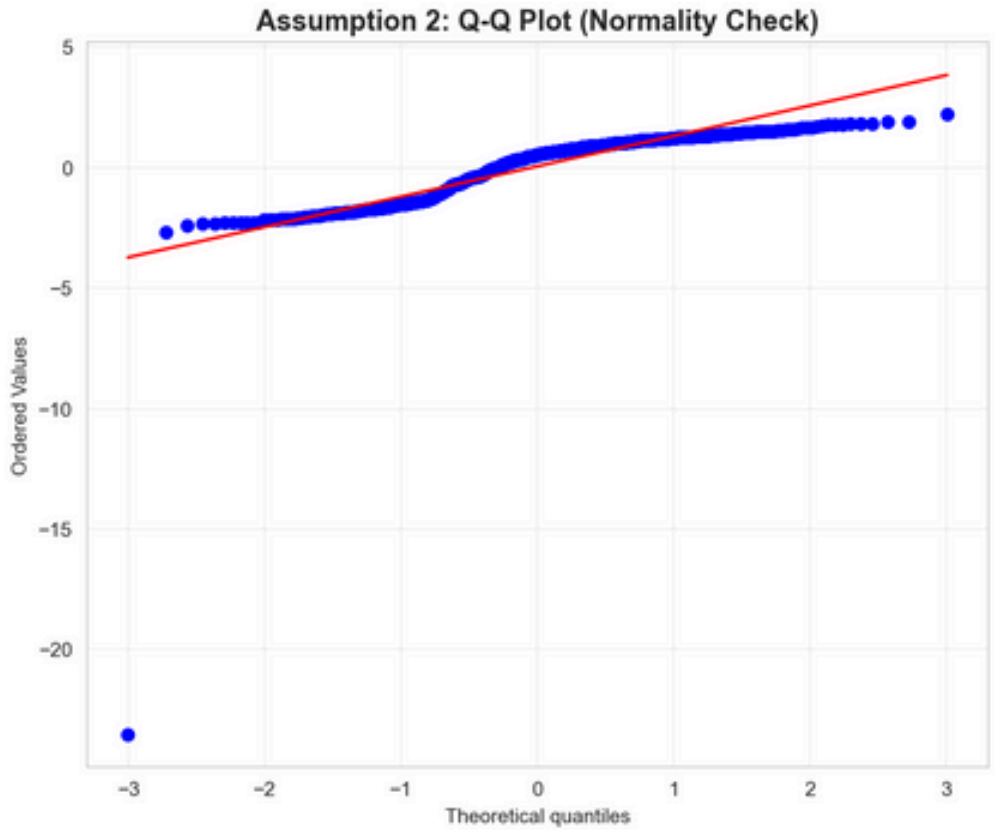
Let e_t denote the residuals of the AR(1) model.

$$H_0 : e_t \sim \mathcal{N}(0, \sigma^2)$$
$$H_1 : e_t \not\sim \mathcal{N}(0, \sigma^2)$$

*Shapiro–Wilk test produces a test statistic of **0.6465** with a p-value approximately equal to 0. Such a small p-value indicates overwhelming evidence against the null hypothesis.*

→ *Therefore, **H_0 is rejected** (p-value < 0.05), and the residuals of the AR(1) model can not be regarded as normally distributed.*

Normality: FAIL



Q-Q plot shows one extreme outlier

3. **Check the assumptions of AR(1) model: zero mean and constant variance of errors (by residuals vs. time plot); normality (by Q–Q plot or Shapiro–Wilk test); no autocorrelation (by ACF of residuals or Ljung–Box test).**

Assumption 3: No Autocorrelation of Errors

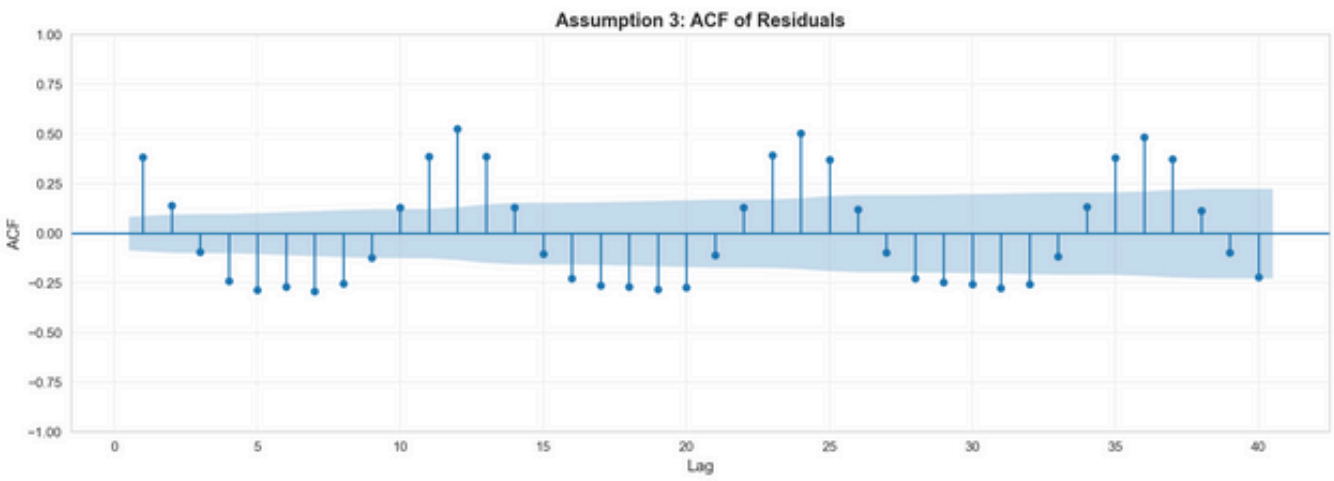
The ACF plot of the residuals shows several autocorrelation coefficients that exceed the 95% confidence bounds at multiple lags, indicating systematic dependence remaining in the residuals.

Let e_t denote the residuals of the AR(1) model.

$$H_0 : \rho_e(1) = \rho_e(2) = \cdots = \rho_e(h) = 0 \quad \text{(no autocorrelation up to lag } h\text{)}$$
$$H_1 : \exists k \leq h \text{ such that } \rho_e(k) \neq 0$$

*Ljung–Box test results at lags **1, 12, 24** and **36**, all of which yield extremely large test statistics and p-values effectively equal to zero. These results provide overwhelming evidence against the null hypothesis of no autocorrelation.*

→ *Therefore, **H_0** (p-value<0.05) **is rejected**, implying that the residuals of the AR(1) model are not white noise and still exhibit significant autocorrelation.*



ACF shows significant spikes at lags 1, 12, 24, 36 (seasonal pattern)

No Autocorrelation: FAIL

3. **Check the assumptions of AR(1) model: zero mean and constant variance of errors (by residuals vs. time plot); normality (by Q–Q plot or Shapiro–Wilk test); no autocorrelation (by ACF of residuals or Ljung–Box test).**

Summary

1. Zero Mean & Constant Variance: **PASS**
2. Normality: **FAIL**
3. No Autocorrelation: **FAIL**

Overall Conclusion

AR(1) on original data is **inadequate** due to non-stationarity, non-normality, and strong seasonal autocorrelation. The model cannot capture trend and seasonality in CO2 data.