

정상성 검정 (Stationarity Test)

정상성

- 시계열 데이터의 통계적 특성(평균, 분산, 공분산 등)이 시간의 흐름에 따라 변하지 않고 일정하게 유지되는 성질을 의미
- 많은 시계열 모델(특히 ARIMA 계열)은 데이터가 정상성을 만족한다고 가정
→ 모델링 전에 시계열 데이터의 정상성 여부를 확인하는 것이 매우 중요
- 비정상 시계열은 ****차분(Differencing)이나 변환(Transformation)****을 통해 정상 시계열로 변환
 - 평균이 일정하지 않으면 **차분**(안될 경우 반복) / 분산이 일정하지 않으면 **변환**

정상성 검정

- 시계열 데이터가 정상성을 만족하는지 통계적으로 확인하는 방법
- 가장 널리 사용되는 검정 방법은 **확장 디키-풀러 검정 (Augmented Dickey-Fuller, ADF Test)**

확장 디키-풀러 검정 (Augmented Dickey-Fuller, ADF Test)

- 시계열 데이터에 ****단위근(Unit Root)****이 존재하는지 여부를 검정하여 정상성을 판단
- 단위근이 존재하면 비정상 시계열이고, 단위근이 존재하지 않으면 정상 시계열
- 가설:
 - **귀무가설 (\$H_0\$)**: 시계열에 단위근이 존재한다 (즉, 비정상 시계열이다).
 - **대립가설 (\$H_1\$)**: 시계열에 단위근이 존재하지 않는다 (즉, 정상 시계열이다).
- 판단 기준:
 - **p-value**
 - **p-value**가 유의수준(e.g. 0.05)보다 작으면 귀무가설을 기각하고 대립가설을 채택
→ 시계열이 정상성을 만족한다고 판단
 - **ADF 통계량**: ADF 통계량이 임계값(Critical Value)보다 작으면 귀무가설을 기각

적용 가능한 상황

- ARIMA, SARIMA 등 정상성을 가정하는 시계열 모델을 적용하기 전에 데이터의 정상성 여부를 확인해야 할 때.
- 시계열 데이터에 추세나 계절성이 존재하는지 통계적으로 확인하고 싶을 때.
- 비정상 시계열을 정상 시계열로 변환(차분 등)한 후, 변환된 데이터가 정상성을 만족하는지 재확인할 때.

구현 방법

`statsmodels` 라이브러리의 `adfuller` 함수를 사용합니다.

주의사항

- **p-value 해석**: p-value가 유의수준보다 크다고 해서 반드시 비정상 시계열이라고 단정할 수는 없습니다. 데이터의 특성과 시각적 분석을 함께 고려해야 합니다.
- **차분 횟수**: ADF 검정 결과 비정상 시계열로 판단되면, 차분을 수행하여 정상 시계열로 변환합니다. 몇 번의 차분을 해야 하는지는 ACF/PACF 플롯이나 반복적인 ADF 검정을 통해 결정합니다.

- **데이터 길이:** ADF 검정은 충분히 긴 시계열 데이터에 대해 신뢰할 수 있는 결과를 제공합니다. 데이터 길이가 너무 짧으면 검정력이 낮아질 수 있습니다.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.stattools import adfuller

# 1. 비정상 시계열 데이터 생성 (추세 포함)
np.random.seed(42)
n_samples = 100
data_trend = np.cumsum(np.random.randn(n_samples) * 0.5 + 0.1) + np.linspace(0, 10, n_samples)
ts_non_stationary = pd.Series(data_trend, index=pd.date_range(start='2020-01-01', periods=n_samples, freq='D'))

# ADF 검정 함수
def adf_test(series):
    result = adfuller(series, autolag='AIC')
    print('ADF Statistic: %f' % result[0])
    print('p-value: %f' % result[1])
    print('Critical Values:')
    for key, value in result[4].items():
        print('\t%s: %.3f' % (key, value))

    if result[1] <= 0.05:
        print("귀무가설 기각 (p-value <= 0.05): 시계열은 정상성을 만족합니다.")
    else:
        print("귀무가설 채택 (p-value > 0.05): 시계열은 비정상 시계열입니다.")

# 2. ADF 검정 수행 (비정상 시계열)
plt.figure(figsize=(10, 4))
plt.plot(ts_non_stationary)
plt.title('Non-Stationary Time Series (with Trend)')
plt.show()

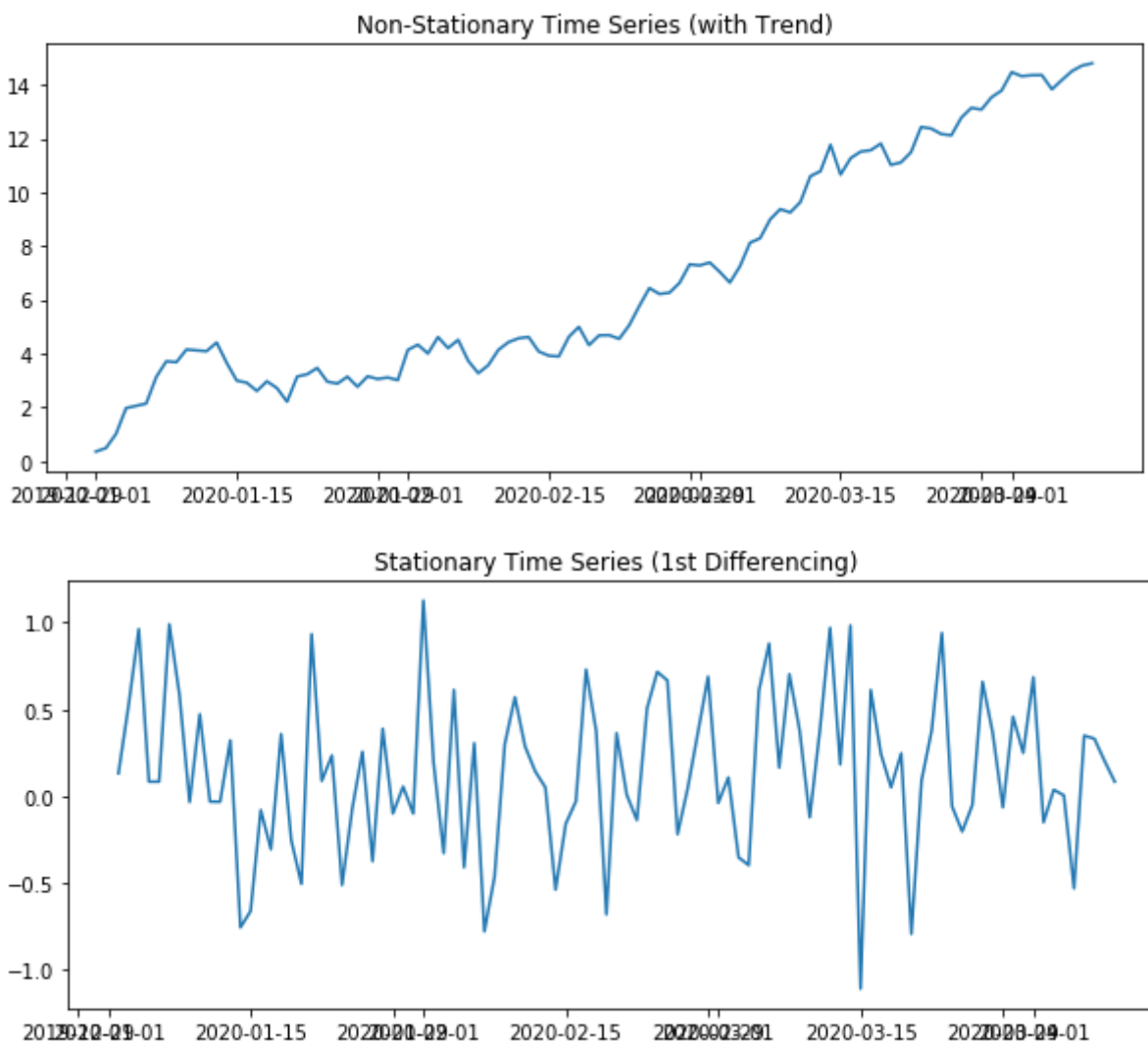
print("--- 비정상 시계열에 대한 ADF 검정 결과 ---")
adf_test(ts_non_stationary)
...
ADF Statistic: 0.027081
p-value: 0.960715
Critical Values:
    1%: -3.498
    5%: -2.891
   10%: -2.583
귀무가설 채택 (p-value > 0.05): 시계열은 비정상 시계열입니다.
...

# 3. 1차 차분 수행하여 정상 시계열로 변환
ts_stationary = ts_non_stationary.diff().dropna()

plt.figure(figsize=(10, 4))
```

```
plt.plot(ts_stationary)
plt.title('Stationary Time Series (1st Differencing)')
plt.show()

# 4. ADF 검정 수행 (정상 시계열)
print("\n--- 1차 차분 후 시계열에 대한 ADF 검정 결과 ---")
adf_test(ts_stationary)
...
ADF Statistic: -10.008881
p-value: 0.000000
Critical Values:
    1%: -3.499
    5%: -2.892
    10%: -2.583
귀무가설 기각 (p-value <= 0.05): 시계열은 정상성을 만족합니다.
...
```



결과 해석 방법

- **ADF Statistic:** 검정 통계량입니다. 이 값이 음수이고, 절댓값이 클수록 정상성에 가깝습니다.
- **p-value:** 귀무가설(단위근 존재, 비정상)을 기각할지 여부를 결정하는 값입니다. 일반적으로 0.05보다 작으면 정상 시계열로 판단합니다.

- **Critical Values:** 1%, 5%, 10% 유의수준에서의 임계값입니다. ADF 통계량이 이 임계값보다 작으면 해당 유의수준에서 귀무가설을 기각할 수 있습니다.

장단점 및 대안

- **장점:**
 - 시계열 데이터의 정상성 여부를 통계적으로 객관적인 기준으로 판단할 수 있습니다.
 - ARIMA 모델링 전 데이터 전처리(차분)의 필요성을 결정하는 데 중요한 지표가 됩니다.
- **단점:**
 - p-value가 유의수준보다 크다고 해서 반드시 비정상 시계열이라고 단정할 수는 없으며, 다른 시각적 분석(시계열 플롯, ACF/PACF)과 함께 고려해야 합니다.
 - 데이터의 길이가 짧거나, 구조적 변화가 있는 경우 검정력이 낮아질 수 있습니다.
- **대안:**
 - **KPSS 검정 (Kwiatkowski-Phillips-Schmidt-Shin Test):** ADF 검정과 가설이 반대입니다.
 - 귀무가설 (H_0): 시계열이 정상성을 만족한다.
 - 대립가설 (H_1): 시계열에 단위근이 존재한다 (즉, 비정상 시계열이다).
 - ADF와 KPSS 검정을 함께 사용하여 정상성 여부를 더 확실하게 판단할 수 있습니다.
 - **시각적 분석:** 시계열 플롯, 이동 평균/분산 플롯, ACF/PACF 플롯 등을 통해 시계열의 추세, 계절성, 자기상관 등을 육안으로 확인하는 것도 중요합니다.