
(보안데이터분석) 연습문제_10

---

1. 다음 중 횡단적 분석(Cross-sectional Analysis)과 종단적 분석(Longitudinal Analysis)에 대한 비교 설명으로 **가장 적절하지 않은** 것은 무엇인가?
    1. **횡단적 분석:** 특정 시점(단일 시점)에 여러 대상(개인, 집단 등)으로부터 데이터를 수집하여 변수들 간의 관계나 집단 간의 차이를 파악하는 데 중점을 둔다.
    2. **종단적 분석:** 동일한 대상으로부터 시간의 흐름에 따라 반복적으로 데이터를 수집하여 대상의 변화, 발달 과정 또는 인과 관계를 추적하는 데 중점을 둔다.
    3. **장점 비교:** 횡단적 분석은 비교적 짧은 시간과 적은 비용으로 수행할 수 있는 반면, 종단적 분석은 시간과 비용이 많이 소요될 수 있지만, 변화의 추이와 인과 관계 파악에 유리하다.
    4. **데이터 독립성:** 횡단적 분석 데이터는 일반적으로 각 관측치가 서로 독립적이라고 가정하는 반면, 종단적 분석 데이터는 시간적 종속성(temporal dependency)을 가진다.
    5. **적용 분야:** 횡단적 분석은 시간에 따른 변화를 직접적으로 측정하는 데 매우 효과적인 반면, 종단적 분석은 특정 시점의 광범위한 현황 파악에만 적합하다. ✅
2. 다음 중 시계열 데이터(Time Series Data)가 일반적인 횡단면(Cross-sectional) 데이터나 다른 유형의 데이터와 구별되는 **가장 핵심적인 특징**은 무엇인가?
    1. 데이터의 양이 매우 방대하여 빅데이터 분석 기법이 필수적으로 요구된다.
    2. 데이터 포인트 간에 측정 단위가 통일되어 있지 않아 정규화가 반드시 필요하다.
    3. 데이터의 각 관측치가 시간 순서에 따라 의존성을 가지며, 이 순서가 분석에 매우 중요하다. ✅
    4. 모든 데이터 포인트가 서로 독립적이며, 순서가 바뀌어도 분석 결과에 영향을 미치지 않는다.
    5. 주로 범주형(categorical) 변수로 구성되어 있어 더미 변수화가 필수적이다.
3. 시계열 데이터의 '분해 기법(Decomposition Techniques)'에 대한 설명으로 **가장 적절하지 않은** 것은 무엇인가?

   1. 시계열 데이터를 추세(Trend), 계절성(Seasonality), 불규칙 요인(Irregular/Residual) 등의 구성 요소로 분리하여 각 요소의 특성을 이해하고 예측 모델링을 돕는 기법이다.
   2. 분해 모델은 크게 **가법 모델(Additive Model)** 과 **승법 모델(Multiplicative Model)** 로 나뉘며, 이는 계절성 변동의 폭이 추세의 크기에 비례하는지 여부에 따라 선택된다.
   3. **가법 모델**은 시계열 = 추세 + 계절성 + 불규칙 요인의 형태로 표현되며, 계절성 변동의 크기가 시간에 따라 일정하다고 가정할 때 적합하다.
   4. **승법 모델**은 시계열 = 추세 × 계절성 × 불규칙 요인의 형태로 표현되며, 계절성 변동의 크기가 추세의 수준에 비례하여 증가하거나 감소할 때 적합하다.
   5. 시계열 분해 기법은 데이터의 패턴을 명확히 파악하는 데 유용하지만, 분해된 각 요소를 재결합하여 미래 값을 예측하는 데는 전혀 사용되지 않는다. ✅
4. 시계열 분석에서 '화이트 노이즈(White Noise)'의 특징과 중요성에 대한 설명으로 **가장 적절하지 않은** 것은 무엇인가?
   1. **특징:** 화이트 노이즈는 평균이 0이고 분산이 일정하며, 서로 다른 시점의 관측치들 사이에 어떠한 자기 상관성(autocorrelation)도 없는 무작위적인 시계열이다.
   2. **예측 불가능성:** 화이트 노이즈는 본질적으로 어떠한 예측 가능한 패턴도 가지고 있지 않으므로, 미래 값을 예측하는 것이 불가능하다.
   3. **모델링의 목표:** 시계열 모델링의 궁극적인 목표 중 하나는 원본 시계열에서 추세, 계절성 등의 모든 계통적인 패턴을 제거하여 잔차(residuals)를 화이트 노이즈에 가깝게 만드는 것이다.
   4. **모델 성능 지표:** 모델 학습 후 잔차가 화이트 노이즈 특성을 보인다면, 이는 모델이 시계열의 예측 가능한 부분을 성공적으로 포착했음을 의미하는 좋은 지표이다.
   5. **예측의 핵심:** 화이트 노이즈는 시계열 데이터의 예측 가능한 부분의 핵심 구성 요소이므로, 예측 정확도를 높이기 위해 이를 잘 모델링해야 한다. ✅
5. 시계열 데이터에서 '자기 상관성(Autocorrelation)'에 대한 설명으로 **가장 적절하지 않은** 것은 무엇인가?
   1. 자기 상관성은 시계열 데이터의 현재 관측치와 과거의 특정 시점의 관측치 사이의 선형적 관계를 나타내는 통계적 측정값이다.
   2. 시계열 데이터가 자기 상관성을 가진다는 것은 데이터의 각 관측치가 시간 순서에 따라 서로 독립적이지 않고, 과거 값이 현재 값에 영향을 미친다는 것을 의미한다.
   3. 자기 상관 함수(ACF: Autocorrelation Function)는 다양한 시차(lag)에서의 자기 상관 계수를 그래프로 나타내어, 시계열의 패턴과 주기성을 파악하는 데 사용된다.
   4. 부분 자기 상관 함수(PACF: Partial Autocorrelation Function)는 다른 시차의 영향을 제거하고, 특정 시차에서의 순수한 자

기 상관 관계를 측정하여 모델의 차수를 결정하는 데 도움을 준다.
    5. 화이트 노이즈(White Noise) 시계열은 강한 자기 상관성을 가지므로, 예측 모델이 화이트 노이즈 특성을 잔차로 남기면 모델이 데이터를 잘 설명하지 못했음을 의미한다.✅
6. 시계열 분석에서 '단위근(Unit Root)'에 대한 설명으로 **가장 적절하지 않은** 것은 무엇인가?
    1. 단위근은 시계열 데이터가 비정상성(non-stationarity)을 가지게 하는 원인 중 하나로, 시계열의 분산이나 평균이 시간에 따라 변하는 특징을 갖게 한다.
    2. 단위근을 포함하는 시계열은 일반적으로 현재 시점의 값이 이전 시점의 값에 강하게 의존하며, 충격이 발생하면 그 영향이 시간이 지나도 소멸되지 않고 지속되는 경향이 있다.
    3. 단위근 검정(Unit Root Test, 예: Dickey-Fuller Test)은 시계열에 단위근이 존재하는지 여부를 통계적으로 확인하기 위해 사용되는 방법이다.
    4. 단위근이 존재하는 시계열은 주로 차분(differencing)을 통해 정상성(stationarity)을 확보할 수 있으며, 이는 시계열 분석의 전처리 과정에서 매우 중요하다.
    5. 단위근은 시계열 데이터가 항상 일정한 평균과 분산을 가지며, 시간의 흐름에 따라 통계적 특성이 변하지 않는다는 것을 의미한다. ✅
7. 시계열 분석에서 '자기회귀모델(Autoregressive Model, AR Model)'에 대한 설명으로 **가장 적절하지 않은** 것은 무엇인가?
    1. 자기회귀모델은 시계열의 현재 값이 이전 시점의 자기 자신의 값들에 선형적으로 의존한다는 가정을 기반으로 한다.
    2. AR(p) 모델에서 'p'는 모델의 차수(order)를 나타내며, 현재 값을 예측하는 데 사용되는 과거 시점의 관측치 개수를 의미한다.
    3. 자기회귀모델은 시계열 데이터가 정상성(stationarity)을 가질 때 가장 효과적으로 적용될 수 있으며, 비정상성 시계열에는 직접 적용하기 어렵다.
    4. 자기회귀모델의 계수(coefficient)는 부분 자기 상관 함수(PACF) 그래프를 통해 유의미한 시차(lag)를 파악하여 결정하는 데 도움을 받을 수 있다.
    5. 자기회귀모델은 현재 값을 예측하기 위해 과거 시점의 '예측 오차(error term)'에 가중치를 부여하여 사용한다. ✅
8. 시계열 데이터의 '평활화 기법(Smoothing Techniques)'에 대한 설명과 그 방법에 대한 진술 중 **가장 적절하지 않은** 것은 무엇인가?
    1. 평활화 기법은 시계열 데이터에 포함된 불규칙한 변동(노이즈)을 제거하여, 데이터의 장기적인 추세(Trend)나 계절성(Seasonality)과 같은 근본적인 패턴을 명확하게 드러내는 것을 목적으로 한다.
    2. **이동평균(Moving Average)** 은 가장 기본적인 평활화 기법 중 하나로, 특정 기간 동안의 관측값들의 평균을 계산하여 시계열을 부드럽게 만들며, 기간이 길어질수록 평활화 정도가 강해진다.

3. **지수 평활(Exponential Smoothing)** 은 과거 관측치에 지수적으로 감소하는 가중치를 부여하여 현재 시점의 평활값을 계산하는 방식으로, 최근 관측치에 더 큰 중요도를 부여한다.
4. Holt-Winters 지수 평활법은 추세(Trend)와 계절성(Seasonality) 요인을 모두 고려하여 예측하며, 계절성 패턴의 크기가 시간에 따라 변하는 경우 승법 모델(Multiplicative Model)을 사용할 수 있다.
5. 시계열 평활화는 데이터의 정상성(Stationarity)을 확보하는 주된 방법으로 사용되며, 단위근(Unit Root) 문제를 해결하기 위해 주로 평활화 기법을 적용한다. ✅ 차분(Differencing)이 사용된다.

- 다음은 python의 statsmodels 패키지를 사용한 시계열 분해 예시 코드이다.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.seasonal import seasonal_decompose

# 1. 시계열 데이터 생성 (예시: 가상의 월별 판매 데이터)
np.random.seed(42)
n_years = 5
n_months = n_years * 12
index = pd.date_range(start='2020-01-01', periods=n_months, freq='MS')

# 추세 (선형 증가)
trend = np.linspace(100, 150, n_months)

# 계절성 (월별 패턴)
seasonal = 10 * np.sin(np.linspace(0, 2 * np.pi * 1, n_months) * 12 / (2 *
np.pi)) + \
           5 * np.cos(np.linspace(0, 2 * np.pi * 2, n_months) * 12 / (2 *
np.pi))

# 불규칙 요인 (노이즈)
np.random.seed(7)
irregular = np.random.normal(0, 5, n_months)

# 가법 모델 (Additive Model)로 시계열 생성
data = trend + seasonal + irregular
ts_data = pd.Series(data, index=index)

# 2. 시계열 분해 수행 (가법 모델)
decomposition = seasonal_decompose(ts_data, model='additive', period=12) #
월별 계절성이므로 period=12

# 3. 분해 결과 시각화
plt.figure(figsize=(12, 8))
plt.subplot(411)
plt.plot(ts_data, label='Original')
```

```
plt.legend(loc='upper left')
plt.subplot(412)
plt.plot(decomposition.trend, label='Trend')
plt.legend(loc='upper left')
plt.subplot(413)
plt.plot(decomposition.seasonal, label='Seasonal')
plt.legend(loc='upper left')
plt.subplot(414)
plt.plot(decomposition.resid, label='Residual')
plt.legend(loc='upper left')
plt.tight_layout()
plt.show()

# 4. 승법 모델 (Multiplicative Model)로 분해 수행 (예시)
# data_multiplicative = trend * (1 + seasonal/100) * (1 + irregular/100) #
승법 모델에 맞는 데이터 생성 예시
# ts_data_multiplicative = pd.Series(data_multiplicative, index=index)
# decomposition_multi = seasonal_decompose(ts_data_multiplicative,
model='multiplicative', period=12)
# plt.figure(figsize=(12, 8))
# decomposition_multi.plot()
# plt.tight_layout()
# plt.show()
```

9. 위 코드를 사용하여 `statsmodels` 의 `seasonal_decompose` 함수를 이용한 시계열 분해에 대한 설명으로 **가장 적절하지 않은** 것은 무엇인가?
    1. `seasonal_decompose` 함수는 시계열 데이터를 추세(Trend), 계절성(Seasonal), 불규칙 요인(Residual)의 세 가지 구성 요소로 분해한다.
    2. `model` 인자는 시계열이 'additive'(가법) 모델인지 'multiplicative'(승법) 모델인지를 지정하며, 이는 계절성 변동의 방식에 따라 선택된다.
    3. `period` 인자는 시계열 데이터의 계절성 주기를 명시하며, 위 코드에서 월별 데이터이므로 12로 설정할 수 있다.
    4. 분해 결과는 `decomposition.trend` , `decomposition.seasonal` , `decomposition.resid` 와 같은 속성을 통해 각 구성 요소에 접근할 수 있다.
    5. `seasonal_decompose` 함수는 내부적으로 이동평균(Moving Average) 방법을 사용하여 추세와 계절성 요인을 추출하며, 이 과정에서 항상 시계열 데이터의 정상성(Stationarity)을 자동으로 확보해준다. ✅ 정상성을 '자동으로 항상 확보'해주는 것은 아니다.
10. 위 코드로 수행된 시계열 분해 결과와 그 활용에 대한 설명으로 **가장 적절하지 않은** 것은 무엇인가?
    1. 분해된 추세(Trend) 컴포넌트는 시계열 데이터의 장기적인 상승 또는 하강 경향을 파악하는 데 사용될 수 있다.

2. 분해된 계절성(Seasonal) 컴포넌트는 매년 또는 매 주기마다 반복되는 특정 패턴(예: 여름철 판매 증가)을 시각화하고 이해하는 데 유용하다.
3. 분해된 불규칙 요인(Residual)은 모델이 설명하지 못한 무작위적인 변동을 나타내며, 이 잔차가 화이트 노이즈에 가까울수록 모델이 시계열의 패턴을 잘 포착했음을 의미한다.
4. 분해된 각 컴포넌트(추세, 계절성, 불규칙 요인)는 독립적으로 미래 값을 예측한 후, 이들을 합치거나 곱하여 원본 시계열의 미래를 예측하는 데 활용될 수 있다.
5. 가법 모델(Additive Model)로 가법 모델은 계절성 변동의 크기가 추세의 변화에 비례하여 증가할 때 적합하며, 이 경우 `model='multiplicative'` 옵션을 사용해야 한다. ✅