

그레인저 인과성 검정 보고서

그레인저 인과성 검정 이해(귀무가설 및 대립가설)

Reference : <https://velog.io/@euisuk-chung/개념정리-시계열-분석-인과관계-분석>

목적 : 매크로팀에서 제공받은 매크로 지표들이, 진짜로 ETF 포트폴리오 수익률을 예측 / 설명 할 수 있는가? → 먼저 인과관계가 존재하는지 확인하고자 함.

- 상관관계와 인과관계의 주요 차이점

특성	상관관계	인과관계
방향성	없음	있음
원인-결과	불명확	명확
제3요인 영향	가능	배제
예측력	제한	강력

개념

그랜저 인과관계(Granger causality)는 1969년 클라이브 그랜저가 제안한 개념으로, 시계열 데이터에서의 예측 가능성을 기반으로 한 인과관계를 의미합니다.

그랜저 인과관계의 정의

(Def) X의 과거 정보가 Y의 미래 값을 예측하는 데 통계적으로 유의미한 도움을 준다면, "X는 Y를 그랜저 인과한다"고 말합니다.

P 이는 수학적으로 다음과 같이 표현할 수 있습니다:

$$P(Y(t+1)|Y(t), Y(t-1), \dots, X(t), X(t-1), \dots) \neq P(Y(t+1)|Y(t), Y(t-1), \dots))$$

- 만약 X(독립)가 Y(종속)을 그레인저 인과한다면, X의 과거 값들을 포함하여 예측한 Y의 미래 값(좌변)이 Y의 과거 값들만으로 예측한 경우(우변)과 다를 것.
→ X의 정보는 Y의 예측에 유의미한 영향을 준다.
- 💡 적용 : 통계검정을 위해서는 보통 선형회귀 모델을 사용하여 두 경우(X 포함 vs. X 미포함)의 예측오차를 비교함.
이외 F-검정이나 다른 통계적 방법을 사용해서 두 모델 간 차이가 유의미한지 판단.
- 주의점 : 예측 가능성에 기반한 인과관계이며, **진정한 인과관계를 완벽히 증명하지는 않음**. 숨겨진 변수나 공통원인 등의 가능성을 항상 고려해야 함.

전제조건

- 시간 선후관계 : 과거의 사건은 현재의 사건을 유발할 수 있으나, 미래의 사건은 현재의 사건을 유발할 수 없다.
- 정상성(stationary) : 평균과 분산이 시간에 따라 일정해야 한다.(AR 조건)
- 입력시차 : 예상되는 시차만큼의 과거 데이터를 모두 입력변수로 사용해야 한다.
- 최종시차 : 적절한 시차를 선택해야 한다. 보통 연 환산 빈도의 2 ~ 3배까지 고려한다.

모형의 검정

- 귀무가설 설정, H_0 = "X는 Y를 그레인저 인과하지 않는다."
- 제한 모델과 비제한 모델 구축

- 제한 모델 : $Y(t) = \alpha + \beta_1 Y(t-1) + \beta_2 Y(t-2) + \dots + \beta_p Y(t-p) + \varepsilon$
- 비제한 모델 : $Y(t) = \alpha + \beta_1 Y(t-1) + \dots + \beta_p Y(t-p) + \gamma_1 X(t-1) + \dots + \gamma_p X(t-p) + \varepsilon$

3. F-통계량 계산

$$F = \frac{(RSS_{restricted} - RSS_{unrestricted})/m}{RSS_{unrestricted}/(n-k)}$$

- $RSS_{restricted}$: 제한된 모델(과거의 Y만 사용)의 잔차 제곱합
- $RSS_{unrestricted}$: 확장된 모델(X와 Y 모두 사용)의 잔차 제곱합
- m : 추가된 파라미터 수 (시차의 수)
- n : 관측치의 수
- k : 확장된 모델의 총 파라미터 수

4. p-value 확인 및 결론 도출

p-value가 0.05 이하일 경우 귀무가설(즉, X는 Y를 그레인저 인과하지 않는다)을 기각할 수 있으며, 이는 X가 Y를 그레인저 인과한다는 것을 의미.

모형의 응용(추후 공부)

- VAR 활용
<https://songseungwon.tistory.com/133>
- ARIMA 활용(현업에서 LSTM보다 우세한 경우 존재)

타겟 변수(ETF, Y)와 9/21 매크로팀 제공 변수(factor = feature, X)와의 그레인저 인과성 검정 코드 작성

주의사항 : 각 타겟 ETF와 각 팩터 간의 관계를 독립적으로 검증하는 방식인 일대일 비교 방법을 사용해야 함. ex) 11개의 타겟 ETF와 20여 개의 각 팩터 간의 관계를 독립적으로 검증한다. / **양방향으로 모두 검증해야 한다.**

▼ 이유

1. 서로 다른 금융 팩터를 그레인저 인과관계 검증에 추가할 수 있나요?

네, 서로 다른 여러 개의 금융 팩터(예: Mkt-Rf, CPI지수, SMB, RMA 등)를 그레인저 인과관계 검증에 추가하여 **동시에 검증**할 수 있습니다. 이는 **다변량 분석**으로, 여러 독립 변수가 하나의 종속 변수에 대해 영향을 미치는지를 검증하는 방식입니다. 이 경우 각 금융 팩터가 ****해당 ETF의 주가(Y)****에 미치는 영향을 확인할 수 있습니다.

2. 이 과정은 VAR(Vector AutoRegression) 모델로 가능한가요?

네, 이 과정은 **VAR(벡터 자기회귀) 모델**을 통해 가능합니다. VAR 모델은 **다수의 시계열 변수**를 동시에 모델링하며, 각 변수의 과거 값이 다른 변수들의 미래 값을 예측하는 데 얼마나 기여하는지를 분석합니다. 즉, VAR 모델을 사용하면 여러 금융 팩터(Mkt-Rf, CPI, SMB 등)가 ETF 주가에 미치는 영향을 **동시에 분석**할 수 있습니다.

3. 일반 AR 모델로는 이 과정을 검증할 수 없고, Y와 X1, Y와 X2의 인과성을 따로따로 검증해야 하나요?

맞습니다. 일반 **AR(자기회귀) 모델**은 **단일 시계열 변수**만을 다루기 때문에, Y와 각각의 X 변수를 **개별적으로** 분석해야 합니다. 즉, **AR 모델**로는 각 금융 팩터와 ETF 주가 간의 **일대일 인과성**을 검증하는 것밖에 할 수 없습니다. 따라서 여러 금융 팩터가 동시에 영향을 미치는 경우에는, AR 모델 대신 ****다변량 모델(예: VAR 모델)****을 사용해야 합니다.

프로토타입

```

import numpy as np
import pandas as pd
from statsmodels.tsa.stattools import grangercausalitytests as test

# 검정 시작 ~ 종료일 설정
start_date = "2000-01-01"
end_date = "2024-09-01"

# 샘플 데이터 생성: 10개의 타겟 ETF와 20개의 팩터 -> 실제 데이터로 대체 필요
np.random.seed(42)
n = 100 # 각 시계열의 길이
num_etfs = 11
num_factors = 20
# 가상의 ETF 및 팩터 데이터 생성 -> 실제 데이터로 대체 필요
etf_data = {f'ETF_{i+1}': np.random.randn(n) for i in range(num_etfs)}
factor_data = {f'Factor_{j+1}': np.random.randn(n) for j in range(num_factors)}

# 데이터프레임 생성
etf_df = pd.DataFrame(etf_data)
factor_df = pd.DataFrame(factor_data)

# 결과를 저장할 데이터프레임 초기화
results = []

# 일대일 비교를 위한 반복문
for etf in etf_df.columns:
    for factor in factor_df.columns:
        # X -> Y 검정
        test_result_XY = test(pd.DataFrame({'X': etf_df[etf], 'Y': factor_df[factor]}), maxlag = 1)
        p_value_XY = test_result_XY[1][0]['ssr_ftest'][1] # 시차 1의 p-value

        # Y -> X 검정
        test_result_YX = test(pd.DataFrame({'X': factor_df[factor], 'Y': etf_df[etf]}), maxlag = 1)
        p_value_YX = test_result_YX[1][0]['ssr_ftest'][1] # 시차 1의 p-value

        # 인과성 여부 판단
        granger_causality = (p_value_XY < 0.05) and (p_value_YX < 0.05)

        # 결과 저장
        results.append({
            'ETF': etf,
            'Factor': factor,
            'p-value (X->Y)': p_value_XY,
            'p-value (Y->X)': p_value_YX,
            'Granger Causality': granger_causality
        })

# 결과를 데이터프레임으로 변환
results_df = pd.DataFrame(results)

# 결과를 CSV 파일로 저장
results_df.to_csv('granger_causality_results.csv', index=False)

```

9/21

1. 제공받은 매크로변수 확인

- **중요도 1.xlsx → 2개**

df1 = 미국 신규 실업수당청구건수(주별 데이터 1967.02.18 ~ 2024.09.14),

df2 = 미국 GDP Now GDP Forecast(일별 데이터 2011.08.25 ~ 2024.09.18)

- **중요도 2.xlsx → 7개**

df3 = 미국실질GDP연율전기대비(**분기별** 데이터 1947년 2분기 ~ 2024년 2분기),

df4 = 미국ISM제조업지수,

df5 = 미국ISM비제조업지수,

df6 = 미국CPI원지수,

(상기 3개 월별.월말 데이터 1997.07 ~ 2024.08)

df7 = 미국 BEI인플레이션채권-10년-기대인플레이션(일별 데이터 1997.08.31 ~ 2024.09.19),

df8 = WTI(일별 데이터 1983.03.30 ~ 2024.09.19),

df9 = 금값(일별 데이터 1994.08.11 ~ 2024.09.19)

- **중요도 3.xlsx → 4개**

df10 = 미국실업률_미국SA,

df11 = PCE 가격지수 전년대비,

df12 = Core PCE 가격지수 전년대비,

df13 = 비농업부문고용자수

(모두 월별.월말 데이터 1959.01 ~ 2024.07)

2. 매크로변수 정성분석

- 시간상 GPT 활용하였음.

▼ 분석결과

중요도 1.xlsx - 2개 변수

1. 미국 신규 실업수당 청구 건수 (주별 데이터):

- **고용 시장 상태**를 실시간으로 반영하는 지표입니다. 신규 실업수당 청구 건수가 증가하면 고용 시장이 악화되고, 경제 성장에 부정적인 영향을 미칠 수 있습니다. 반면 감소는 고용 시장이 강세임을 의미합니다. 이는 경기 침체와 경기 회복을 예측하는 선행 지표로 활용됩니다.

2. 미국 GDP Now GDP Forecast (일별 데이터):

- 미국 애틀랜타 연준에서 제공하는 실시간 GDP 추정치로, **향후 성장률**을 실시간으로 예측하는 데 유용합니다. 금융인들은 이를 통해 기업 실적 전망, 금리 변동 예측, 투자 전략을 세우는 데 활용할 수 있습니다.

중요도 2.xlsx - 7개 변수

1. 미국 실질 GDP 연율 전기 대비 (분기별 데이터):

- **경제 성장을 평가하는 핵심 지표**입니다. 실질 GDP 성장률이 높다면 경제가 확장되고 있음을 의미하며, 낮거나 음수라면 경기 침체를 나타냅니다. 중앙은행의 금리 정책과 긴밀히 연결되어 있어, 금융시장에서 큰 주목을 받습니다.
2. **미국 ISM 제조업 지수 (월별):**
 - **제조업 경기의 건강 상태**를 나타내는 지표입니다. 50을 기준으로 경기 확장(50 이상)과 축소(50 이하)를 판단하며, 제조업의 생산활동, 주문량 등을 반영합니다. 이는 제조업 관련 기업 주식이나 원자재 가격에 직접적인 영향을 미칩니다.
 3. **미국 ISM 비제조업 지수 (월별):**
 - **서비스업 경기**를 측정하는 지표입니다. 서비스업이 경제에서 차지하는 비중이 높기 때문에 제조업 지수만큼 중요합니다. 금융 시장에서는 서비스업에 대한 경기 흐름을 예측하고, 소비와 관련된 산업에 대한 투자 결정을 내리는 데 도움을 줍니다.
 4. **미국 CPI 원지수 (월별):**
 - **소비자 물가 변동**을 나타내는 지표로, 인플레이션의 척도입니다. 물가가 상승하면 중앙은행이 금리 인상을 통해 이를 억제할 가능성이 높아지며, 이는 채권 수익률과 통화 가치에 큰 영향을 미칩니다. CPI는 주식 및 채권 시장의 주요 리스크 요인입니다.
 5. **미국 BEI 인플레이션 기대 (일별):**
 - **10년 만기 인플레이션 기대치**로, 시장이 장기적으로 예상하는 인플레이션 수준을 보여줍니다. 금리 및 인플레이션 연계 채권을 거래하는 투자자들에게 중요한 지표로, 미래의 금리와 인플레이션 환경을 예측하는 데 활용됩니다.
 6. **WTI 유가 (일별):**
 - **세계 경제와 밀접하게 연결된 원유 가격**입니다. 유가 상승은 원가 상승과 인플레이션 압박을 의미하며, 이는 운송, 에너지, 소비재 기업에 큰 영향을 미칩니다. 유가가 하락하면 경제 성장에 긍정적인 요인으로 작용할 수 있습니다.
 7. **금값 (일별):**
 - **안전자산으로서의 금 가격**은 시장의 리스크 선호도와 밀접하게 연결되어 있습니다. 경제 불확실성이 커지면 금값이 상승하고, 경제가 안정되면 금값은 하락하는 경향이 있습니다. 투자자들은 경기 변동 및 시장 리스크에 대응하여 금 투자를 고려합니다.

중요도 3.xlsx - 4개 변수

1. **미국 실업률 (월별):**
 - **고용 시장의 상태**를 나타내는 핵심 지표입니다. 실업률이 높으면 경제가 침체되어 있음을 의미하고, 실업률이 낮으면 경제가 호황일 가능성이 큼니다. 고용 상태는 소비 지출과 직접적인 상관관계가 있어, 실물 경제에 중요한 영향을 미칩니다.
2. **PCE 가격 지수 전년 대비 (월별):**
 - **개인 소비 지출을 반영한 물가지수**로, 중앙은행이 금리 결정을 할 때 가장 중요하게 참고하는 물가 지표입니다. PCE 가격 지수가 상승하면 인플레이션을 억제하기 위해 금리 인상 가능성이 높아집니다.
3. **Core PCE 가격 지수 전년 대비 (월별):**
 - **에너지와 식품을 제외한 PCE 물가지수**로, 변동성이 적어 **중앙은행이 물가 변동성을 관찰**하는데 중점을 두는 지표입니다. 금리 인상 및 통화 정책 결정에 있어 중요하게 활용됩니다.
4. **비농업부문 고용자수 (월별):**
 - **미국 고용 시장의 상태**를 반영하는 매우 중요한 지표입니다. 비농업 부문은 미국 경제에서 대부분의 고용을 차지하기 때문에, 비농업 부문 고용자 수 변화는 경기와 소비 경향을 예측하는 데 중요한 역할을 합니다.

3. 변수명 지정

원본 이름(중요도)	변수명(영어)
df1. 미국 신규 실업수당청구건수(1)	initial_jobless_claim
df2. 미국 GDP Now GDP Forecast(1)	gdp_now_forecast
df3. 미국 실질 GDP 연율 전기 대비(2)	real_gdp_qoq
df4. 미국 ISM 제조업지수(2)	ism_manufacturing
df5. 미국 ISM 비제조업지수(2)	ism_nonmanufacturing
df6. 미국 CPI 원지수(2)	cpi
df7. 미국 BEI 10년 기대 인플레이션(2)	bei_breakeven_inflation
df8. WTI 유가(2)	wti_oil_price
df9. 금값(2)	gold_price

원본 이름(중요도)	변수명(영어)
df10. 미국 실업률(3)	unemployment_rate
df11. PCE 가격 지수 전년 대비(3)	pce_price_index_yoy
df12. Core PCE 가격 지수 전년 대비(3)	core_pce_price_index_yoy
df13. 비농업부문 고용자수(3)	nonfarm_payroll

4. 매크로변수 별 xlsx → df 변환(자문필요) / 5. 타겟 ETF(11개) 확인

질문사항

1. 타겟 ETF 11개는 조정종가 월말 데이터 활용하면 되는지?
2. 일별 데이터 / 주별 데이터 / 분기별 데이터(매크로 변수들) → 월말 데이터로 리샘플링하는 방법?

일별 데이터 → 월말 날짜만 매칭시킴

주별 데이터, 분기별 데이터는 어떤 방식으로 리샘플링?

임의처리방식

- 일별/주별 데이터의 경우 월말로 리샘플링
- 분기별 데이터의 경우 월별로 forward fill 적용
- 결측치 forward fill + backward fill 적용

→ 심각한 수준의 데이터 결손 및 노이즈 추가의 원인이 될 것으로 예상. 제대로 된 분석이라 하기 어려움.

6. 그레인저 인과성 검정 진행 / 7. 분석결과 및 평가

전반적인 인과성 패턴

→ 대부분의 ETF와 매크로 변수 간에 그랜저 인과관계가 존재하는 것으로 나타났음. 대부분의 p-value가 0.05 미만으로, 통계적으로 유의미한 관계를 보여준다고 볼 수 있음.

IWM (iShares Russell 2000 ETF)

- 모든 매크로 변수와 강한 인과관계를 보임
- 특히 ISM 비제조업지수(df5)와 BEI 10년 기대 인플레이션(df7)에 대해 매우 낮은 p-value를 나타냄

SPY (SPDR S&P 500 ETF Trust)

- 대부분의 매크로 변수와 강한 인과관계 존재
- GDP Now Forecast(df2)와 ISM 제조업지수(df4)에 특히 민감한 반응을 보임

VTV (Vanguard Value ETF)

- 전반적으로 강한 인과관계를 나타냄
- CPI 원지수(df6)와 BEI 10년 기대 인플레이션(df7)에 대해 특히 낮은 p-value를 보임

VUG (Vanguard Growth ETF)

- 대부분의 매크로 변수와 강한 인과관계를 가짐
- WTI 유가(df8)와의 관계가 특히 두드러짐

MTUM (iShares MSCI USA Momentum Factor ETF)

- 다른 ETF들에 비해 상대적으로 약한 인과관계를 보이거나, 여전히 대부분 유의미함
- 신규 실업수당청구건수(df1)와 **GDP Now Forecast**(df2)에 더 민감한 반응을 보임

QUAL (iShares MSCI USA Quality Factor ETF)

- 대체로 강한 인과관계를 나타냄
- **BEI 10년 기대 인플레이션**(df7)과 **WTI 유가**(df8)에 특히 민감한 반응을 보임

VYMI (Vanguard International High Dividend Yield ETF)

- 대부분의 매크로 변수와 강한 인과관계를 가짐
- **ISM 비제조업지수**(df5)와 **CPI 원지수**(df6)에 대해 특히 낮은 p-value를 나타냄

USMV (iShares MSCI USA Min Vol Factor ETF)

- 다른 ETF들에 비해 상대적으로 약한 인과관계를 보이거나, 여전히 대부분 유의미함
- 신규 실업수당청구건수(df1)와 **미국 실업률**(df10)에 더 민감한 반응을 보임

KBE (SPDR S&P Bank ETF)

- 대부분의 매크로 변수와 강한 인과관계를 나타냄
- **ISM 비제조업지수**(df5)와 **BEI 10년 기대 인플레이션**(df7)에 특히 민감한 반응을 보임

IYK (iShares U.S. Consumer Staples ETF)

- 거의 모든 매크로 변수와 매우 강한 인과관계를 보임
- 특히 신규 실업수당청구건수(df1)와 **ISM 비제조업지수**(df5)에 대해 극히 낮은 p-value를 나타냄

IYC (iShares U.S. Consumer Discretionary ETF)

- 대부분의 매크로 변수와 강한 인과관계를 가짐
- **ISM 비제조업지수**(df5)와 **BEI 10년 기대 인플레이션**(df7)에 특히 민감한 반응을 보임

주요 매크로 변수의 영향력

1. **ISM 비제조업지수**(df5): 거의 모든 ETF에 대해 매우 강한 인과관계를 나타냄
2. **BEI 10년 기대 인플레이션**(df7): 대부분의 ETF에 대해 강한 영향력을 보임
3. **신규 실업수당청구건수**(df1): 많은 ETF에 대해 유의미한 인과관계를 가짐
4. **CPI 원지수**(df6): 여러 ETF에 대해 강한 인과관계를 나타냄

결론

→ 대부분의 ETF가 다양한 매크로 경제 변수와 유의미한 그랜저 인과관계를 가지고 있음을 확인가능.

특히 ISM 비제조업지수, BEI 10년 기대 인플레이션, 신규 실업수당청구건수 등이 ETF 움직임에 중요한 영향을 미치는 것으로 보임.