PROJECT

재무지표를 활용한 투자전략 수립

요인분석과 회귀분석을 중심으로

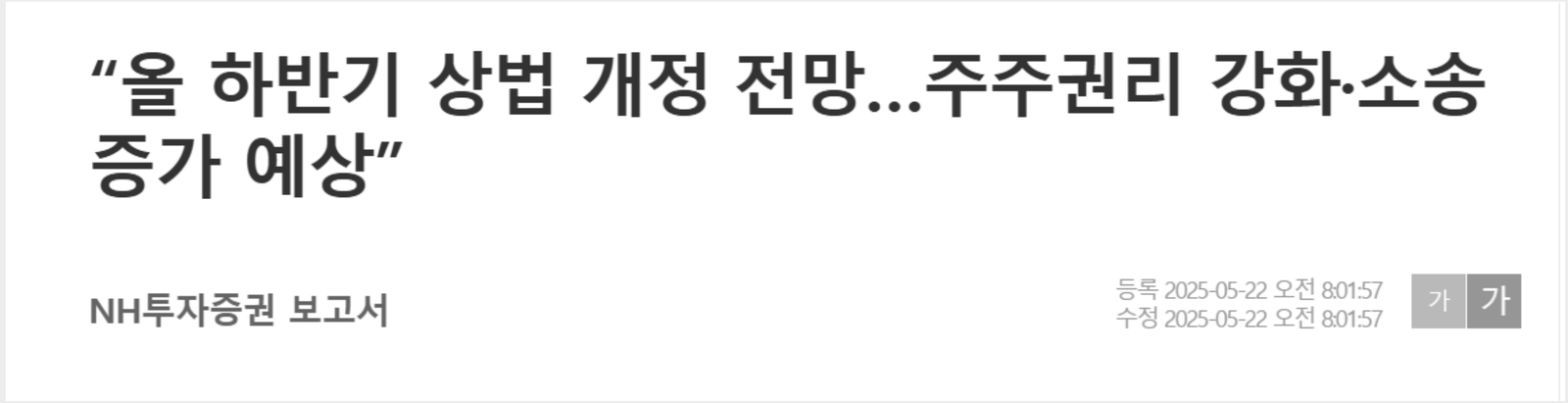
숫자의 마법사들

2025.05.23

✳ Contents

✳ 주제 소개

– 배경 및 선정 이유



“올 하반기 상법 개정 전망…주주권리 강화·소송 증가 예상”

NH투자증권 보고서

등록 2025-05-22 오전 8:01:57
수정 2025-05-22 오전 8:01:57

가 가

‘K 밸류업’이 곧 한국경제 미래…구조개혁으로 펀더멘털 강화해야’ [자본중심 K밸류업⑤]

입력 2024-09-03 16:00

권태성 기자

가

재무제표 속 지표로 수익률을 설명하고, 주식투자 전략을 설계할 수 있을까?

✳ 주제 소개

– 목표 및 기대효과

새로운 지표를 활용한
투자전략 수립/강화

KOSPI200 변화율
이상의 수익률 목표

백테스팅
시각화 서비스

✳ 팀 소개

김보윤 : 팀장, 데이터 전처리, 대시보드 구축

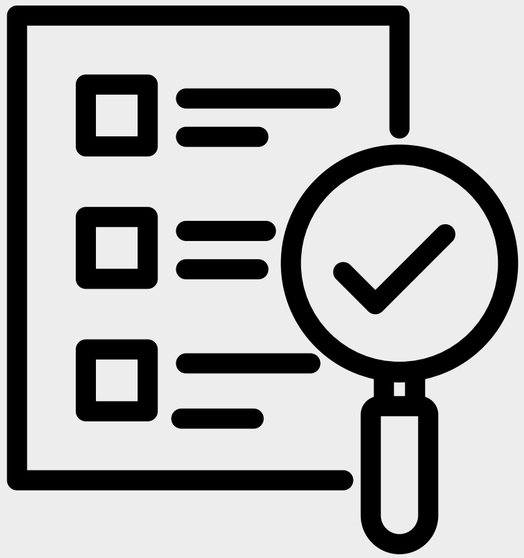
김도현 : 투자지표를 활용한 전략 구성

김성현 : 데이터 구성 및 전처리, 데이터 분석

장규민 : 데이터 분석 모델링 및 검증

정승원 : 투자전략 코드구현

✳ 데이터 전처리



상관분석 / 요인분석

회귀분석

데이터 전처리

✳ 데이터셋 1

- **재무비율지표 : 수치형 데이터**

rows : 2620 columns : 124 (회사명, 주식코드, 연도 제외)

2015년~2024년(TS2000)
코스피200 지수에 포함된 종목(321개 기업)

수익성(39개)	매출액순이익률, 총자본사업익률, 자기자본순이익률, 자본금순이익률, 매출원가 대 매출액비율, 유보율, 자기자본배당율...
활동성(21개)	자본회전률, 자기자본회전률, 자본금회전률, 타인자본회전률, 재고자산회전률, 유동자산회전률, 매출채권회전률...
안정성(38개)	자기자본구성비율, 유동비율, 부채비율, 현금비율, 유보액대비율, 유동부채비율, 차입금비율...
생산성(10개)	기계장비율, 총자본투자효율, 노동소득분배율, 자본분배율...
성장성(16개)	총자본증가율, 유동자산증가율, 재고자산증가율, 자기자본증가율, 매출액증가율, 순이익증가율...

✳ 데이터셋 1

- TS 2000 데이터 전처리

- TS 2000 데이터에서 Null 값 포함 행: 총 40개
- 제조업과 금융업의 재무제표 구조가 상이하여
  - Null이 포함된 제조업 15개 행 우선 제거
- 이후 금융업 종목 전체 제거하여
  - 금융업은 타 업종과 다르게 자본구조, 영업방법, 정부의 규제감독 측면에서 차이가 있어 제외

->분석 대상의 일관성 확보

✳ 데이터셋 2

- 주가 지표 : 수치형 데이터(주식코드 :str)

rows : 3210 columns : 4

2015년~2024년(PyKrx)
코스피200 지수에 포함된 종목(321개 기업)의 연간 수익률

컬럼명	컬럼 소개
주식코드	2015년~2024년 4월 말 기준 코스피200에 포함되었던 종목
연도	2015~2024
종가	매년 4월 말의 종가 데이터(ex. 2024.04.30)
수익률	지난년도 대비 올해의 종가의 변화율 ((당해년도 종가 - 작년 종가) / 당해년도종가)

✳ 데이터셋 2

- PyKrx 데이터 전처리

- PyKrx는 이전 영업일 제공X → 이전 영업일 탐색 함수 작성
  - 4월말 기준 데이터 확보 목적
- 지정 기간 중 KOSPI200에 한 번이라도 포함된 종목 전체 추출
- 종가 기준 수익률 계산 → 수익률 컬럼 생성
- 2025년 종가 데이터가 없는 기업은 상장폐지로 간주하여 제거
- 초기에 종가 컬럼에서 465개 결측 발생
  - 원인: 상장년도 문제, 과거 데이터 부재 등
  - 처리: 결측치 삭제
- 수익률 처리 시
  - ±무한대 → 결측치로 변환
  - 수익률 -1 → 결측치로 변환
- 최종적으로 모든 결측치 제거
  - 분석 목표는 특정 연도 수익률과 재무정보 비교
- 2014년 KOSPI200 종목 데이터는 PyKRX에 없음
  - KRX 정보데이터시스템에서 2014.04.30 데이터 수집하여 병합

✳ 통합데이터셋

- 통합 데이터 : 수치형 데이터

TS2000 재무 데이터				PyKrx 데이터		
EC	ED	EE	EF	EG	EH	EI
부가가치율(IFRS)	노동소득분배율(IFRS)	자본분배율(IFRS)	이윤분배율(IFRS)	시가	종가	수익률
16.9	42.93	57.07	28.86	2577	2540	12.93908
9.09	73.88	26.12	-7.67	2044	2045	-19.4882
12.54	59.44	40.56	17.17	2629	2640	29.09535
12.44	63.85	36.15	4.73	1703	1724	-34.697
9.96	80.37	19.63	-9.85	5510	5070	194.0835
18.99	38.24	61.76	37.38	12050	12350	143.5897
13.96	45.54	54.46	31.5	39500	39500	219.8381
6.16	106.43	-6.43	-36.85	53000	52000	31.64557
14.54	48.47	51.53	25.64	20000	20200	-61.1538
61.39	29.53	70.47	42.99	42150	42750	12.35217
61.25	29.79	70.21	43.84	36350	36800	-13.9181
63.11	31.55	68.45	43.21	29250	29250	-20.5163
58.94	38.98	61.02	35.13	33800	34150	16.75214
59.74	31.99	68.01	37.02	24300	24400	-28.5505
1.52	4509.32	-4409.32	-3840	25550	25150	3.07377
55.58	75.51	24.49	-2.72	26750	26700	6.163022
51.89	53.81	46.19	17.6	18860	19220	-28.015
68.64	37.48	62.52	35.27	15340	15300	-20.3954
75.01	34.49	65.51	42.81	16120	16080	5.098039
35.08	19.66	80.34	20.01	17300	17200	-30.0813
35.04	19.01	80.99	28.16	14600	14450	-15.9884
37.96	18.25	81.75	23.48	14850	15050	4.152249

2015년 회계년도의 재무 데이터(2016년 4월초 공시)
1년 수익률(검증 날짜 : 2015.05 ~ 2016.04)

재무데이터가 4월초에 공개 -> 4월말 리밸런싱 시점 기준

종목코드와 회계년도를 기준으로 내부조인

* 상관분석 / 요인분석

데이터 전처리

회귀분석

상관분석 / 요인분석

✳ 차원 축소

요인분석 (Factor Analysis)	주성분 분석 (Principal Component Analysis)
설명 변수 간 상관관계를 기반으로 공통된 잠재 요인을 추출하는 기법 데이터 차원을 축소하고 핵심 패턴을 파악함	설명 고차원의 데이터의 분산을 최대한 보존하는 방식으로 저차원으로 축소하는 기법
사용 목적 변수 간 구조 파악 및 데이터 차원 축소	사용 목적 요약, 저차원 시각화,노이즈 제거
주요 특징 - 공통 요인 추출 - 변수 요약 및 해석 용이	주요 특징 - 선형 결합, 주성분 간 직교 관계 - 분산 설명력 기준 정렬
활용 방식 주요 요인을 추출한 후, 요인점수를 회귀분석에 활용	활용 방식 데이터 압축을 통한 차원 문제 해결, 다중공선성 문제 해결

✳ 요인 분석

kmo 검정

목적:샘플이 요인분석에 적절한지 측정하는 지표
설명: 변수 간 상관성의 비율을 수치로 나타낸 값 (0~1)
0.6 이상이면 요인분석 가능, 0.8 이상이면 아주 적합

KMO 값: 0.7240144680982331
전체 KMO 값이 약 0.7 정도로 요인분석에 적합함을 확인

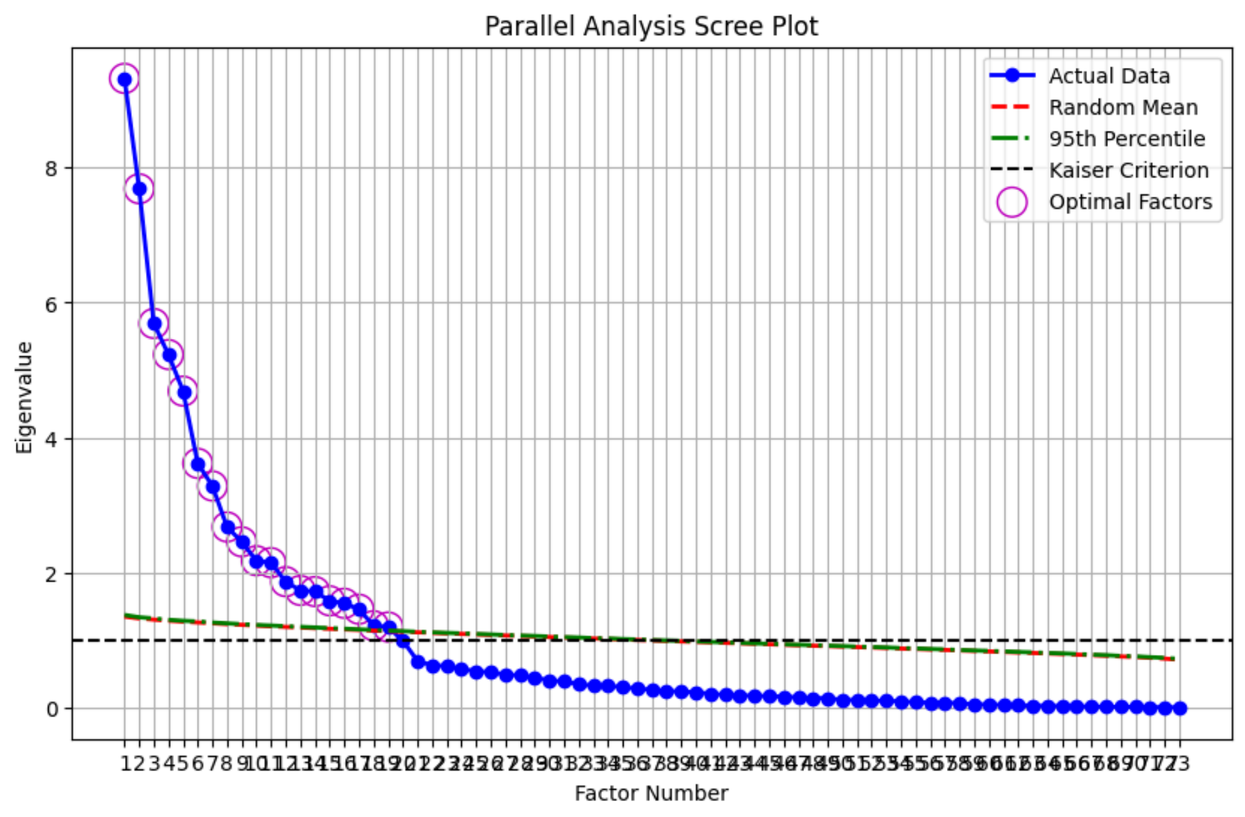
Bartlett 검정

목적: 변수들 간의 상관관계가 요인분석에 적합한지 검정
설명: 상관행렬이 단위행렬이 아닐 확률을 보는 검정

H0 : 상관관계행렬이 단위행렬이다
H1 : 상관관계행렬이 단위행렬이 아니다
(즉 , 변수간 상관관계가 존재한다)

유의수준 1%에서 귀무가설 기각
즉, 변수간 상관관계가 존재하여 요인분석에 적합함을 확인

✳ 분석 결과



피쳐 상관관계 분석 후 |r| ≤ 0.5 변수 제거

스크리 플롯과 카이저 기준선 기반 요인 수 결정

스크리 플롯: 곡선이 완만해지는 지점 활용

카이저 기준: 고유값 ≥ 1인 요인 선택

최종 선택 요인 수: 19개, 누적 분산 비율 : 83.9%

* 상관분석 / 요인분석

데이터 전처리

상관분석 / 요인분석

회귀분석

✳ OLS 회귀분석

OLS 회귀분석 (Ordinary Least Squares)
설명 종속 변수와 독립 변수들 간의 선형 관계를 모델링 회귀계수를 최소제곱법으로 추정하는 방법
사용 목적 변수 간 영향력 분석 및 타겟 변수 설명 모델 구축
주요 특징 - 선형 관계 기반 - 단순 예측 및 인과 분석에 적합
활용 방식 회귀 계수로 변수 영향력 파악, R^2로 예측 성능 평가에 활용

가정	설명
선형성	종속변수와 각 독립변수간 관계가 선형관계
등분산성	모든 관측치에서 오차의 분산이 동일
정규성	잔차는 정규분포를 가짐
독립성	관측치 별 오차항 간 독립

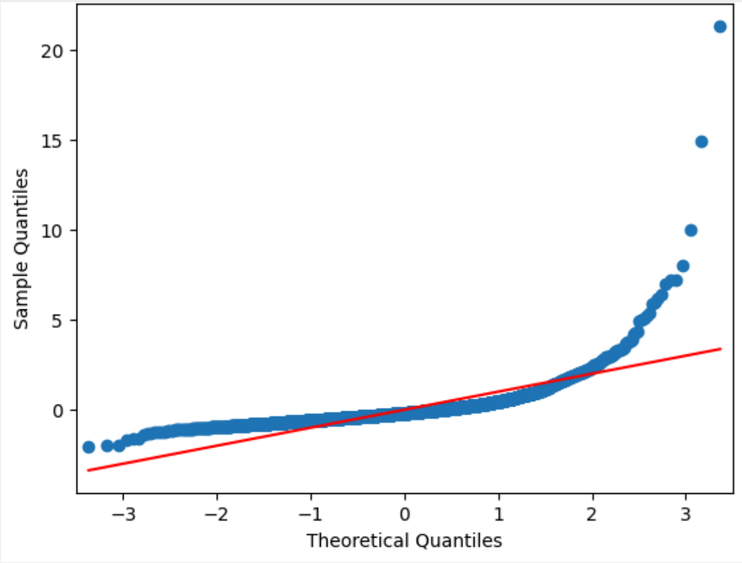
✳ OLS 회귀분석

선형회귀분석

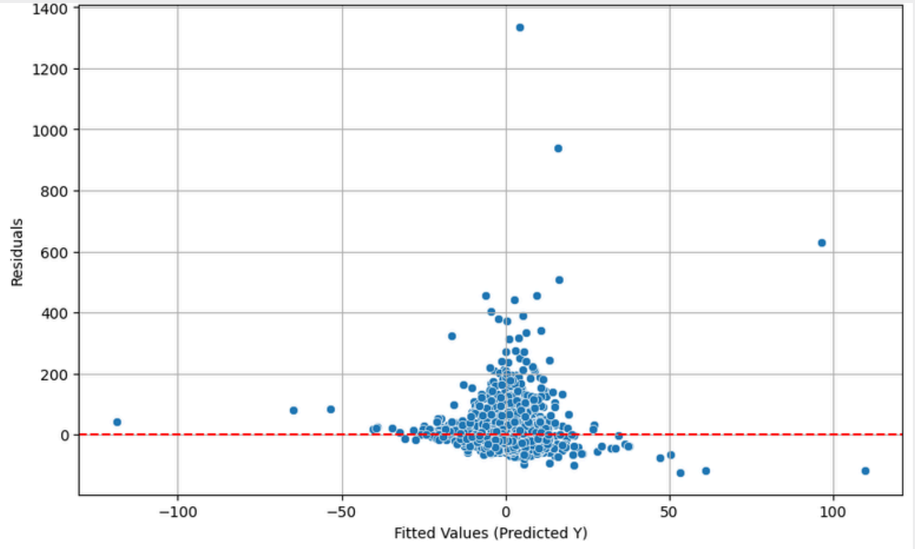
Y의 특성상 잔차가 정규성, 등분산성을 띄지 않음
결정계수 또한 매우 낮음

R-squared	0.017
Adj. R-squared	0.009
F-statistic	2.255
Prob(F-statistic)	0.00115

QQ Plot of Residuals



Residuals vs. Fitted Values



✳ OLS 회귀분석

H0 : 잔차는 정규분포를 따른다
H1 : 잔차는 정규분포를 따르지 않는다

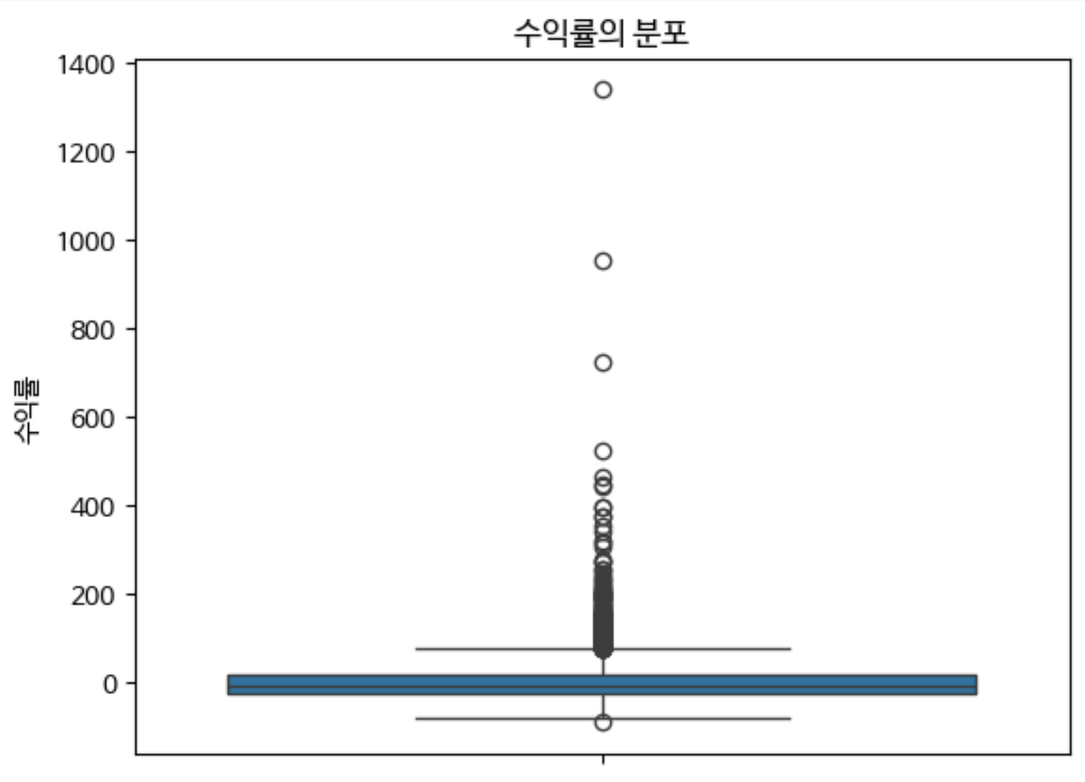
OMnibus	3531.313
Jarque-Bera (JB)	1349361.089

- Levene 검정 통계량: 3.0537
- p-value: 0.0274

→ 잔차의 등분산성이 없다고 판단 (유의수준 5%)

Prob(Omnibus)	0.000	Prob(JB)	0.00
Skew	7.333	Durbin-Watson	2.069
Kurtosis	113.207	Cold. No.	7.09

✳ RLM 회귀분석



RLM 회귀분석 (Robust Linear Model)

설명
이상치의 영향을 최소화하는 회귀 기법
노이즈가 포함된 데이터에서도 안정적인 예측값을 도출함

사용 목적
이상치 영향을 줄여 예측 모델 신뢰도 확보

주요 특징
– 이상치에 강한 회귀계수 추정
– 견고한 모델링 수행

활용 방식
요인 점수를 활용해 성공 확률을 예측하는 회귀모델 구축

✳ 분석 결과

-RLM-
(Robust Linear regression Model)

Robust linear Model Regression Results

Dep. Variable:	수익률	No. Observations:	2620
Model:	RLM	Df Residuals:	2600
Method:	IRLS	Df Model:	19
Norm:	HuberT		
Scale Est.:	mad		
Cov Type:	H1		
Date:	Thu, 22 May 2025		
Time:	07:13:59		
No. Iterations:	16		

Robust linear Model Regression Results

Dep. Variable:	수익률	No. Observations:	2602
Model:	RLM	Df Residuals:	2583
Method:	IRLS	Df Model:	18
Norm:	HuberT		
Scale Est.:	mad		
Cov Type:	H1		
Date:	Fri, 23 May 2025		
Time:	04:37:14		
No. Iterations:	26		

	coef	std err	z	P>|z|	[0.025	0.975]
factor_1	0.0058	0.007	0.871	0.384	-0.007	0.019
factor_2	-0.0144	0.007	-2.174	0.030	-0.027	-0.001
factor_3	0.0143	0.007	2.159	0.031	0.001	0.027
factor_4	0.0074	0.007	1.112	0.266	-0.006	0.020
factor_5	0.0006	0.007	0.096	0.923	-0.012	0.014
factor_6	0.0028	0.007	0.431	0.667	-0.010	0.016
factor_7	-0.0051	0.007	-0.770	0.441	-0.018	0.008
factor_8	0.0799	0.007	12.084	0.000	0.067	0.093
factor_9	0.0222	0.007	3.352	0.001	0.009	0.035
factor_10	-0.0027	0.007	-0.414	0.679	-0.016	0.010
factor_11	0.0106	0.007	1.599	0.110	-0.002	0.024
factor_12	-0.0190	0.007	-2.869	0.004	-0.032	-0.006
factor_13	0.0110	0.007	1.669	0.095	-0.002	0.024
factor_14	0.0424	0.007	6.415	0.000	0.029	0.055
factor_15	0.0007	0.007	0.103	0.918	-0.012	0.014
factor_16	0.0108	0.007	1.639	0.101	-0.002	0.024
factor_17	0.0083	0.007	1.250	0.211	-0.005	0.021
factor_18	-0.0027	0.007	-0.412	0.680	-0.016	0.010
factor_19	-0.0079	0.007	-1.195	0.232	-0.021	0.005

-분산팽창계수-

VIF 지수 결과:

	Variable	VIF
20	factor_20	1.0
14	factor_14	1.0
18	factor_18	1.0
19	factor_19	1.0
17	factor_17	1.0
15	factor_15	1.0
16	factor_16	1.0
13	factor_13	1.0
12	factor_12	1.0
1	factor_1	1.0
11	factor_11	1.0
8	factor_8	1.0
10	factor_10	1.0
9	factor_9	1.0
4	factor_4	1.0
3	factor_3	1.0
2	factor_2	1.0
7	factor_7	1.0
5	factor_5	1.0
6	factor_6	1.0
0	const	1.0

다른 독립변수들과 얼마나 강하게 선형관계를 가지는지 정량화한 지표

VIF = 1
→ 변수들과 상관관계 X

2 ~ 5: 보통 허용 가능한 수준
5 ~ 10: 다중공선성 우려
10: 심각한 다중공선성

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

✳ 프로젝트 일정

- **1주차: 선행연구 분석, ts2000, Pykrx 데이터 수집** — 프로젝트와 유사한 논문 및 관련자료 수집, 10년내 KOSPI200에 속한 종목의 주식, 재무 데이터 수집 및 분석
- **2주차: 데이터 구성 및 전처리** — 결측치 처리 및 변수 정리, 분석용 데이터셋 구축
- **3주차: 투자 전략 구성 및 데이터 분석** — 요인분석 및 회귀분석을 통해 유의미한 팩터 도출
- **4주차: 백테스팅 및 시각화** — 새로운 지표 기반 전략 수립 및 펀드 구성, 성과 시각화

✳ 향후 진행 계획

팩터 전략 고도화

- 유의미한 팩터 기반 포트폴리오 설계

성과 백테스트

- 연도별 수익률 및 벤치마크 수익률과 비교
- 리스크 지표(MDD, 변동성) 포함

시각화 및 해석

- 수익률 추이, 팩터별 기여도 시각화
- 투자 인사이트 도출

Q&A