매매주체별 매수량에 따른 코스피 실증분석

(feat. 다중선형회귀와 로지스틱회귀를 이용한)

주식은 지금 2조

contents

팀 Color

▷ 4명이 모여 만든 시너지 효과

박정호 I 아이디어뱅크

박혜연 I 통계 분석

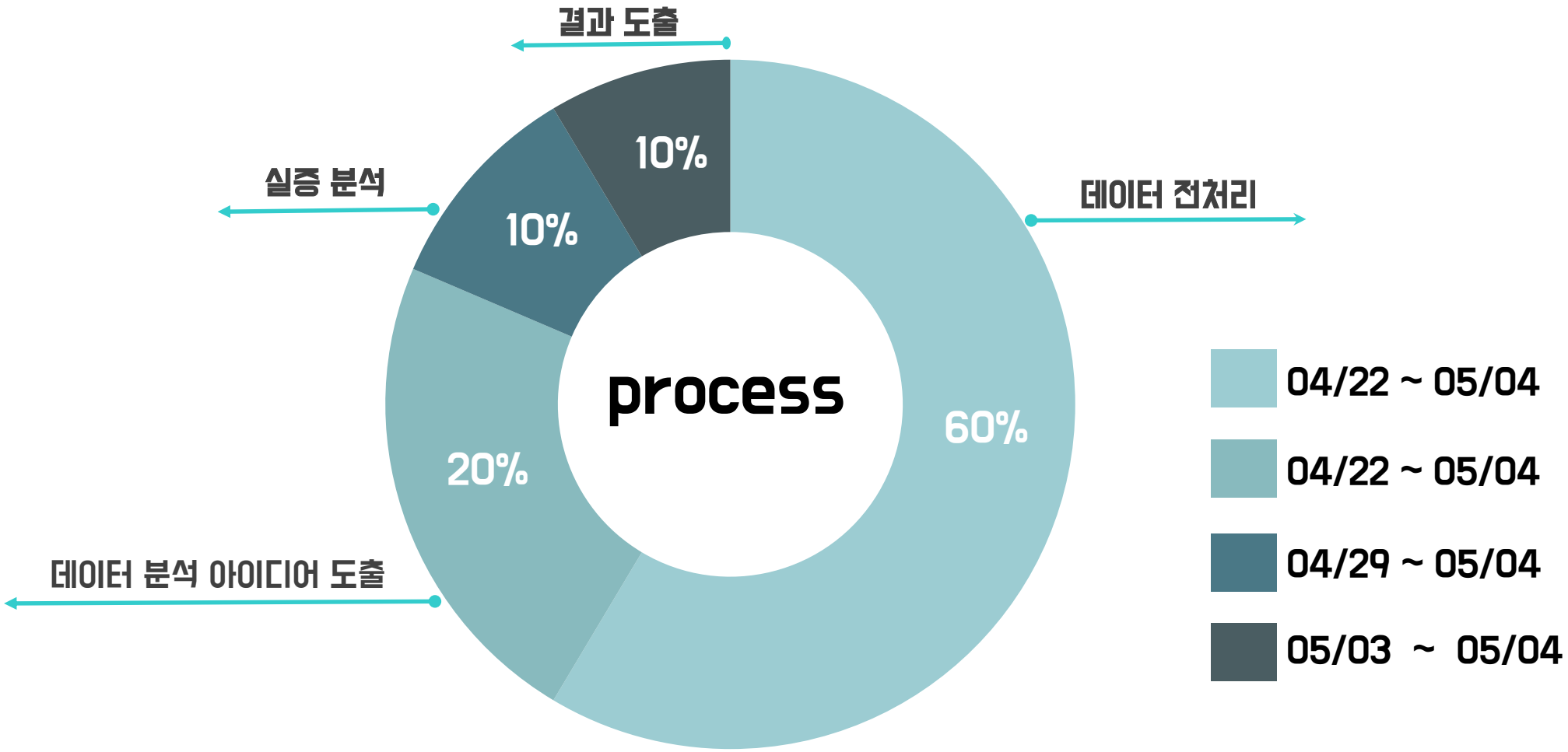
윤새림 I 모델 설계

장동민 I 보완 및 검토

▷ 해커톤을 통해 추구한 우리의 목표

통계의 꽃 "회귀" 완벽 이해 → **모델링 설계**

▷ 해커톤 수행 과정


결과 도출
실증 분석
데이터 전처리
데이터 분석 아이디어 도출
10%
10%
60%
20%
process
04/22 ~ 05/04
04/22 ~ 05/04
04/29 ~ 05/04
05/03 ~ 05/04

기획의도

▷ 매매주체별 매수량에 따라 코스피는 정말 변할까?

“쓸어담은 개미 울고, 내던진 외국인 웃었다 "

“5월엔 팔아라? 외국인 진짜 팔았다… 주가는 주르륵 "

우리나라 유가증권 시장의 순자금 흐름 : 외국인 영향이 가장 큼

실제로 개인/외국인/기관이 코스피에 미치는 영향을 직접 분석하고자 함

2. 분석 방법 소개

▷ 방법1) 다중선형회귀를 이용한 코스피 "등락률" 예측

Y(종속변수) : 2005년 - 2022년 월말 코스피 등락률

X(독립변수) : 개인 , 외국인, 기관 매수량 , 시가총액대비 외국인 코스피 주식 보유 금액 , 동행지수

▷ 방법2) 로지스틱 회귀를 이용한 코스피 "등락" 예측

Y(종속변수) : 2005년 - 2022년 월별 코스피 등락

X(독립변수) : 개인 , 외국인, 기관 매수량 , 시가총액대비 외국인 코스피 주식 보유 금액 , 동행지수

실증분석

▷ 출처 및 수집 과정 소개

	수집한 데이터	출처 및 수집 과정	간격
1	경기선행지수 및 동행지수	ELF라지표 csv 다운로드	월별
2	코스피 종가 및 등락률	Python 내 yfinance . dataReader	일/월별
3	매매주체별 순매수량	네이버 크롤링 / 키움 open API	일/월별
4	시가총액대비 외국인 코스피 주식 보유 금액	한국거래소 csv 다운로드	월별

2. 데이터 전처리

▷ 날짜 변환과 이상치 제거를 중심으로 반복

최종 선택한 변수

✓ Y(종속변수)

일별 코스피 종가 : 정규성 만족 X
일별 코스피 종가 등락률 : 정규성 만족 X
월별 코스피 종가 등락률 : 정규성 만족 O

전처리 과정

1. Yfinance에서 코스피 월말 종가 추출
2. 1번을 바탕으로 전월대비 등락률 계산
3. 이상치 제거 : 2008년 10월 전월대비 등락률

(2008년 금융위기 / 9월15일 리먼브라더스 파산 /9월 월까지 1450였던 코스피가 1000 밑으로)

✓ X(독립변수)

일별 매수량 : 정규성 만족 X
월별 합 매수량 : 정규성 만족 X
일별 이상치 제거 후 월별로 합 : 정규성 만족 O

1. 키움 API에서 매매주체별 일별 매수량 추출
2. 이상치 제거 : 일별 매수량 데이터 중
3. 종속변수와 맞춰주기 위해 월별 합

+ 동행지수와 시가총액대비 외국인 코스피 주식 보유 금액 모두 월별 데이터라 독립변수로 합침

2. 데이터 전처리

▷ 가변수 생성

가변수란? 0 또는 1 값을 갖는 이진수 변수

최종 선택한 변수	전처리 과정
✓ Y(종속변수)	월별 코스피 종가 등락률을 바탕으로
코스피 등락 label	전월 대비 등락률 값이 양수면 1 음수면 0

✓ X(독립변수)

1) 100기준동행지수 label ⋯⋯▸ 동행지수가 100기준 크면 1 작으면 0

2) 동행지수 등락에 따른 label ⋯⋯▸ 전월 대비 등락률 값이 크면 1 작으면 0

2. 데이터 전처리

▷ 최종 데이터 셋

월별

다중선형회귀 종속변수

날짜	년도	월	전월대비 등락률	등락률(%)	등락률label	외국인	기관	개인	시가총액	동행지수	100기준동행지수	동행지수등락률	동행등락label
2005-01-31	2005	1	0.041053	4.105277	1	8538	984	-9523	42.9	99.0	0	-0.100000	0
2005-02-28	2005	2	0.084336	8.433577	1	14654	-8928	-5725	42.7	98.7	0	-0.303030	0
2005-03-31	2005	3	-0.045167	-4.516690	0	-20741	16920	3820	42.2	98.8	0	0.101317	1
2005-04-29	2005	4	-0.056313	-5.631266	0	-3243	955	2291	41.9	98.6	0	-0.202429	0
2005-05-31	2005	5	0.064644	6.464395	1	1048	17254	-18304	41.7	98.7	0	0.101420	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2021-11-30	2021	11	-0.044323	-4.432316	0	11342	-27380	16641	33.0	101.1	1	0.198216	1
2021-12-30	2021	12	0.048834	4.883389	1	15250	15846	-30452	33.5	101.8	1	0.692384	1
2022-01-28	2022	1	-0.105556	-10.555634	0	-389	6687	12674	32.7	102.4	1	0.589391	1
2022-02-28	2022	2	0.013457	1.345673	1	-6207	-3629	-10135	32.4	102.6	1	0.195312	1
2022-03-31	2022	3	0.021662	2.166212	1	-23373	6022	-3604	31.6	102.4	1	-0.194932	0

206 rows × 13 columns

가변수처리 | 로지스틱회귀 종속변수

가변수 처리 | 독립변수

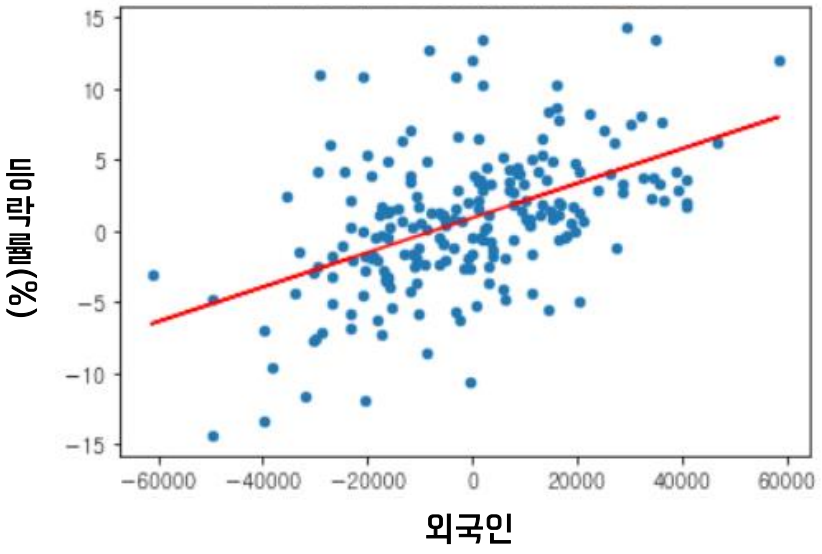
3. 모형설계 - 다중선형회귀를 이용한 등락률 예측

▷ 산점도 및 히트맵을 통한 EDA

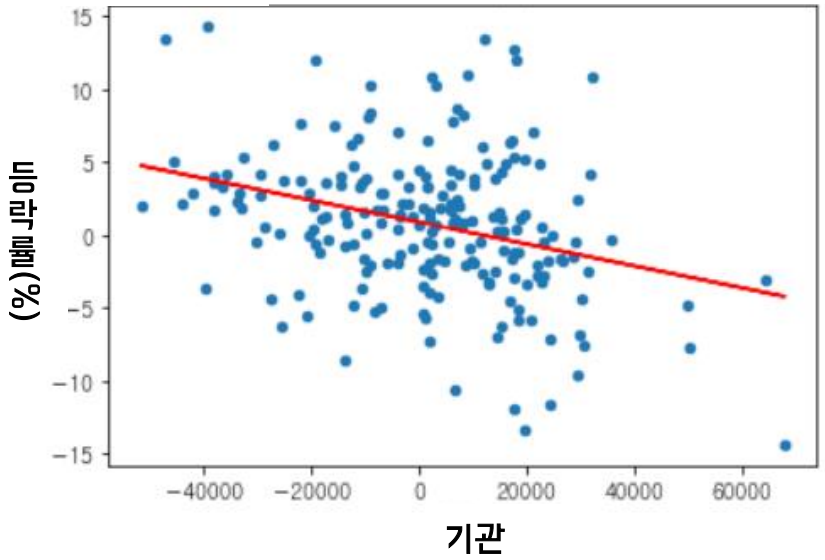
- ✓ 등락률 - 외국인 : + 0.5
- ✓ 등락률 - 기관 : - 0.3
- ✓ 등락률 - 개인 : - 0.4
- ✓ 등락률 - 시가총액 : 0.03
- ✓ 등락률 - 동행지수 : -0.2

산점도 및 히트맵을 통한 EDA

외국인순매수와 등락률(%)사이 산점도



기관순매수와 등락률(%)사이 산점도



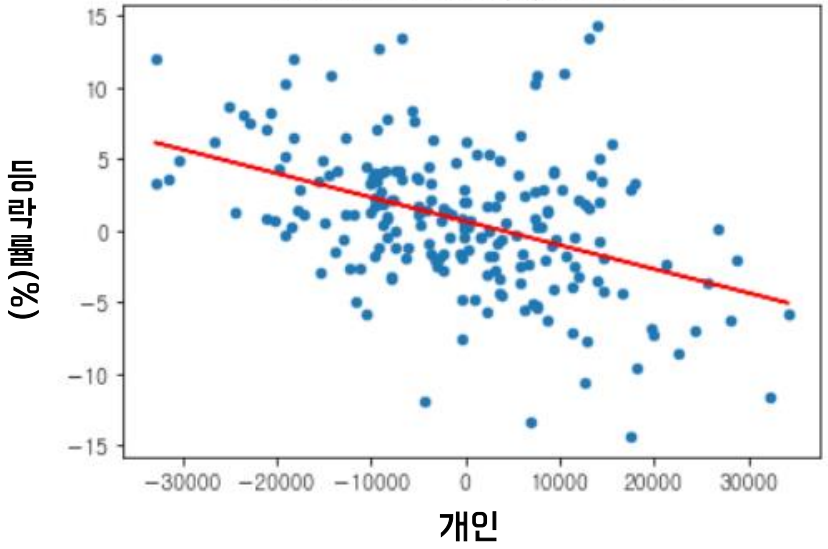
분석 초기 단계 꼭 확인

X-Y 사이 선형 관계 타당성 확인

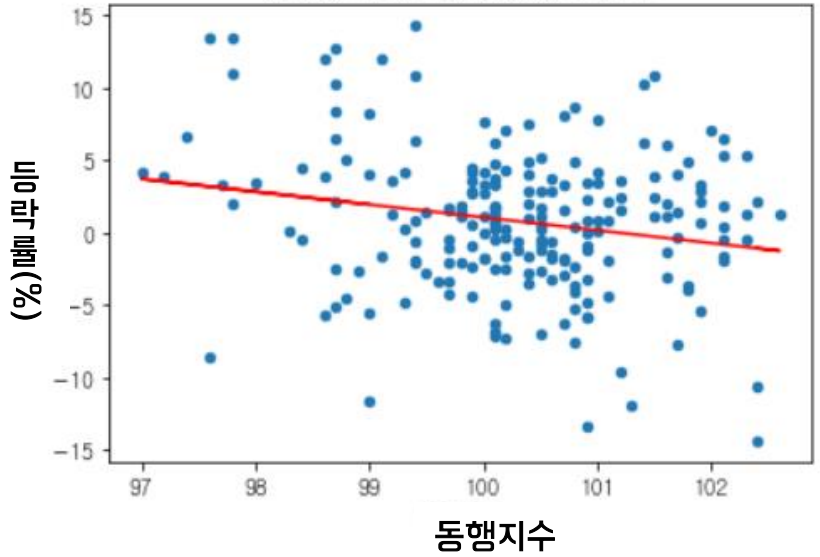
선형성 만족!

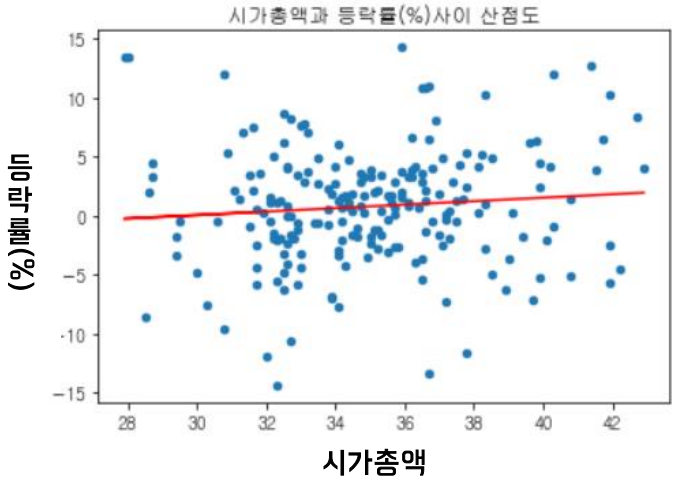
(독립변수와 종속변수 전처리 무한 반복으로 도출)

개인순매수와 등락률(%)사이 산점도



동행지수와 등락률(%)사이 산점도





3. 모형설계 – 다중선형회귀를 이용한 등락률 예측

▷ 다중회귀모형 적합 회귀분석을 수행하기 위한 가설 검정

귀무가설 : 모든 독립변수는 코스피등락률(%) 에 영향을 미치지 않을 것이다.

대립가설: 코스피 등락률(%)에 적어도 하나는 영향을 미칠 것이다.

OLS Regression Results

Dep. Variable:	등락률(%)	R-squared:	0.396
Model:	OLS	Adj. R-squared:	0.381
Method:	Least Squares	F-statistic:	26.04
Date:	Wed, 04 May 2022	Prob (F-statistic):	1.30e-16
Time:	13:25:43	Log-Likelihood:	-454.16
No. Observations:	164	AIC:	918.3
Df Residuals:	159	BIC:	933.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	79.7834	28.138	2.835	0.005	24.211	135.356
외국인순매수	6.657e-05	3.93e-05	1.694	0.092	-1.1e-05	0.000
개인순매수	-0.0001	3.94e-05	-3.726	0.000	-0.000	-6.9e-05
동행지수	-0.7899	0.281	-2.816	0.005	-1.344	-0.236
기관순매수	-1.848e-05	3.55e-05	-0.521	0.603	-8.86e-05	5.17e-05

Omnibus:	11.798	Durbin-Watson:	2.151
Prob(Omnibus):	0.003	Jarque-Bera (JB):	18.082
Skew:	0.391	Prob(JB):	0.000118
Kurtosis:	4.426	Cond. No.	2.53e+06

회귀모형을 적합시키고 F-test 결과를 확인하면 P-value 가 매우 작기 때문에 귀무가설을 기각한다.
즉, 적어도 하나의 독립변수는 종속변수에 유의한 영향을 미친다고 볼 수 있다

3. 모형설계 – 다중선형회귀를 이용한 등락률 예측

▷ 전진 / 후진 / 단계적 선택법으로 독립변수 선택

forward

OLS Regression Results

Dep. Variable:	전월대비 등락률	R-squared:	0.334
Model:	OLS	Adj. R-squared:	0.324
Method:	Least Squares	F-statistic:	33.72
Date:	Tue, 03 May 2022	Prob (F-statistic):	1.03e-17
Time:	21:09:38	Log-Likelihood:	369.50
No. Observations:	206	AIC:	-731.0
Df Residuals:	202	BIC:	-717.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7871	0.256	3.075	0.002	0.282	1.292
외국인순매수	8.441e-07	1.59e-07	5.320	0.000	5.31e-07	1.16e-06
개인순매수	-1.149e-06	2.51e-07	-4.579	0.000	-1.64e-06	-6.54e-07
동행지수	-0.0078	0.003	-3.044	0.003	-0.013	-0.003

Omnibus:	16.435	Durbin-Watson:	1.796
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.103
Skew:	0.552	Prob(JB):	2.61e-05
Kurtosis:	4.113	Cond. No.	1.91e+06

외국인순매수, 개인순매수, 동행지수 | 모두 유의

backward

OLS Regression Results

Dep. Variable:	전월대비 등락률	R-squared:	0.337
Model:	OLS	Adj. R-squared:	0.327
Method:	Least Squares	F-statistic:	34.17
Date:	Tue, 03 May 2022	Prob (F-statistic):	6.59e-18
Time:	21:09:38	Log-Likelihood:	369.95
No. Observations:	206	AIC:	-731.9
Df Residuals:	202	BIC:	-718.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7667	0.256	2.995	0.003	0.262	1.271
개인순매수	-1.841e-06	2.25e-07	-8.186	0.000	-2.28e-06	-1.4e-06
기관순매수	-7.711e-07	1.42e-07	-5.416	0.000	-1.05e-06	-4.9e-07
동행지수	-0.0076	0.003	-2.966	0.003	-0.013	-0.003

Omnibus:	20.448	Durbin-Watson:	1.834
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30.757
Skew:	0.592	Prob(JB):	2.10e-07
Kurtosis:	4.476	Cond. No.	1.85e+06

...

외국인순매수, 기관순매수, 동행지수 | 모두 유의

▷ 전진 / 후진 / 단계적 선택법으로 독립변수 선택

stepwise

OLS Regression Results

Dep. Variable:	전월대비 등락률	R-squared:	0.334
Model:	OLS	Adj. R-squared:	0.324
Method:	Least Squares	F-statistic:	33.72
Date:	Tue, 03 May 2022	Prob (F-statistic):	1.03e-17
Time:	21:09:38	Log-Likelihood:	369.50
No. Observations:	206	AIC:	-731.0
Df Residuals:	202	BIC:	-717.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7871	0.256	3.075	0.002	0.282	1.292
외국인순매수	8.441e-07	1.59e-07	5.32	0.000	5.31e-07	1.16e-06
개인순매수	-1.149e-06	2.51e-07	-4.579	0.000	-1.64e-06	-6.54e-07
동행지수	-0.0078	0.003	-3.044	0.003	-0.013	-0.003

Omnibus:	16.435	Durbin-Watson:	1.796
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.103
Skew:	0.552	Prob(JB):	2.61e-05
Kurtosis:	4.113	Cond. No.	1.91e+06

...

다중선형회귀모형 적합

OLS Regression Results

Dep. Variable:	등락률(%)	R-squared:	0.395
Model:	OLS	Adj. R-squared:	0.383
Method:	Least Squares	F-statistic:	34.79
Date:	Tue, 03 May 2022	Prob (F-statistic):	2.30e-17
Time:	20:49:51	Log-Likelihood:	-454.30
No. Observations:	164	AIC:	916.6
Df Residuals:	160	BIC:	929.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	80.9640	27.983	2.893	0.004	25.701	136.227
외국인순매수	8.497e-05	1.71e-05	4.974	0.000	5.12e-05	0.000
개인순매수	-0.0001	2.74e-05	-4.824	0.000	-0.000	-7.8e-05
동행지수	-0.8017	0.279	-2.874	0.005	-1.353	-0.251

Omnibus:	10.497	Durbin-Watson:	2.145
Prob(Omnibus):	0.005	Jarque-Bera (JB):	15.191
Skew:	0.366	Prob(JB):	0.000503

외국인순매수, 개인순매수, 동행지수 I 모두 유의

R-Squared : 0.395 I 모든변수 → 유의함

전진선택법과 단계선택법의 최종 결과가 같아 단계선택법을 적용한 모형을 최종모형으로

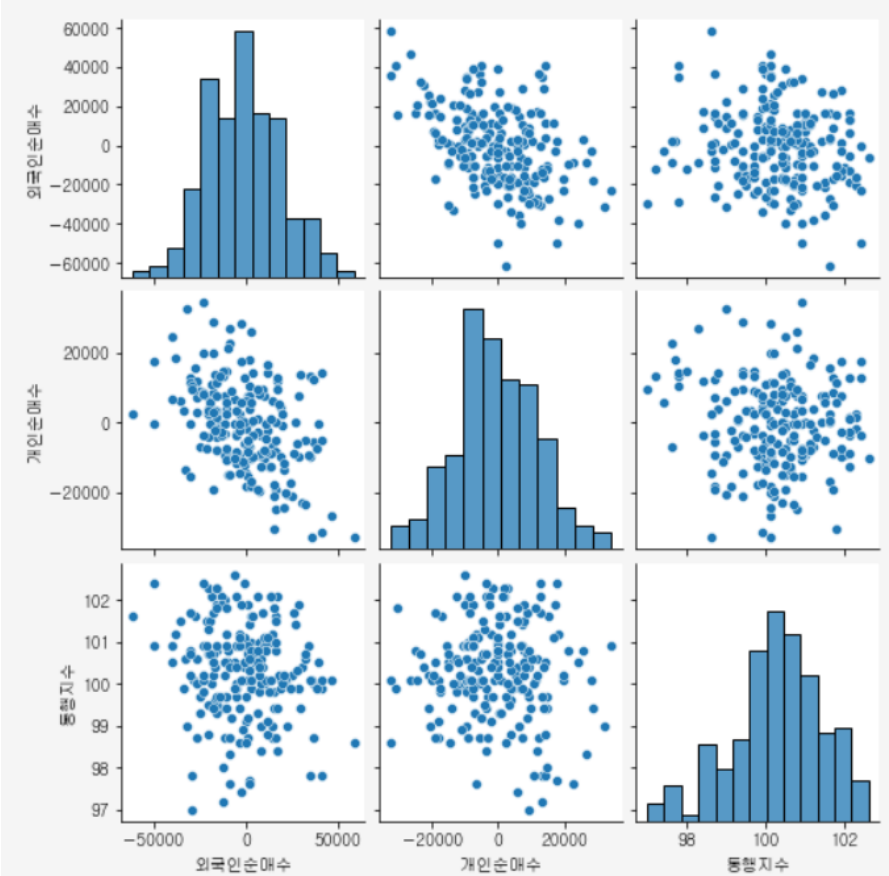
3. 모형설계 – 다중선형회귀를 이용한 등락률 예측

▷ 최종 모형 독립변수들간의 상관관계 검증

```
feature = finish_dummy[['외국인순매수', '개인순매수','동행지수']]
```

```
def vif(data):
    import pandas as pd
    from statsmodels.stats.outliers_influence import variance_inflation_factor

    # VIF 출력을 위한 데이터 프레임 형성
    vif = pd.DataFrame()

    # VIF 값과 각 Feature 이름에 대해 설정
    vif["VIF Factor"] = [variance_inflation_factor(data.values, i) for i in range(len(data.columns))]
    vif["features"] = data.columns

    # VIF 값이 높은 순으로 정렬
    vif = vif.sort_values(by="VIF Factor", ascending=False)
    vif = vif.reset_index().drop(columns='index')

    return vif

vif(feature)
```

	VIF Factor	features
0	1.230653	개인순매수
1	1.228050	외국인순매수
2	1.011263	동행지수



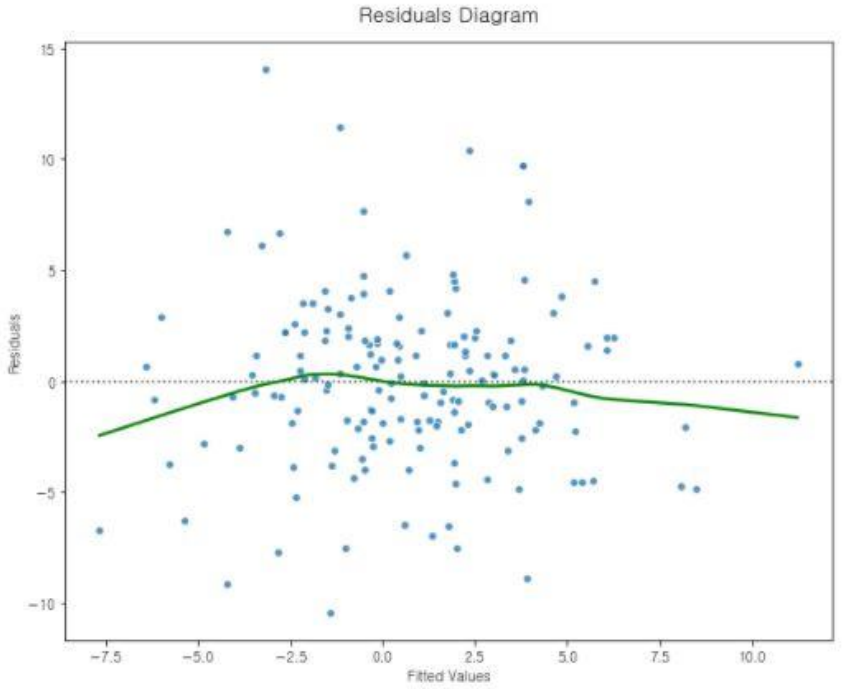
분산 팽창 요인(VIF)를 구했을 때 값이 10보다 작기때문에 **다중공선성이 없다!**

Pair plot을 통해 알아본 다중공선성 시각화

3. 모형설계 - 다중선형회귀를 이용한 등락률 예측

▷ 선형회귀분석의 잔차 3가지가정

1. 오차항의 등분산성

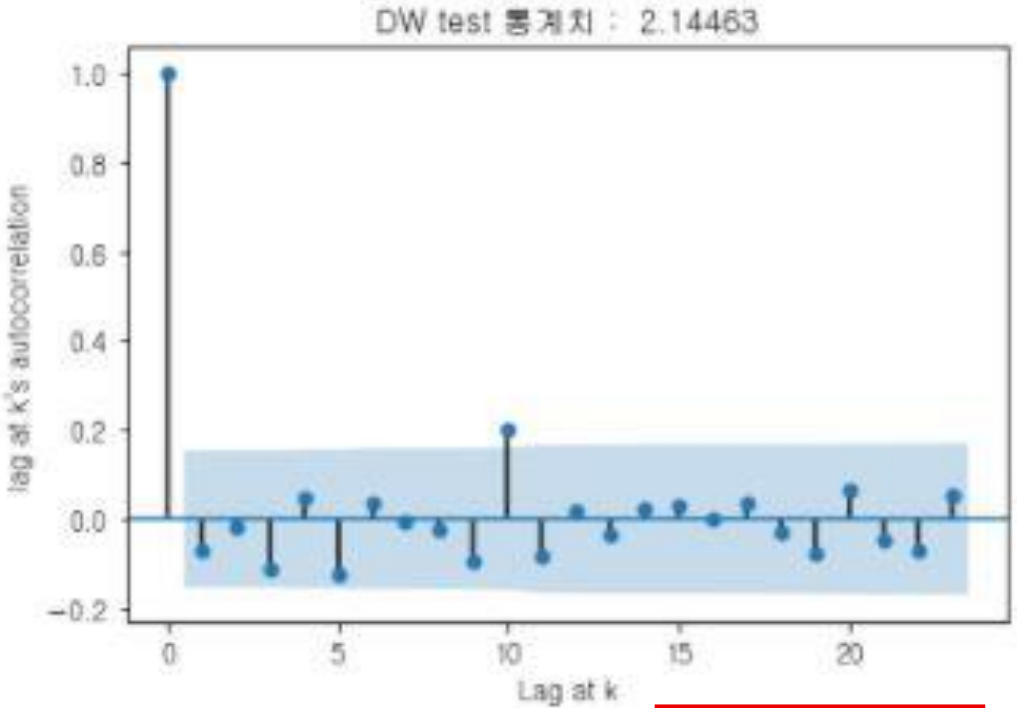


그래프 결과 초록색 실선이 수평선을 그리는것을 시각적으로 확인

오차항이 등분산성을 따른다는 귀무가설을 P- value : 0.271 로 유의수준 5.0%에서 기각하지 못함

등분산성 만족!

2. 오차항의 독립성



DW-test(더빈왓슨 검정) : 2.1446307

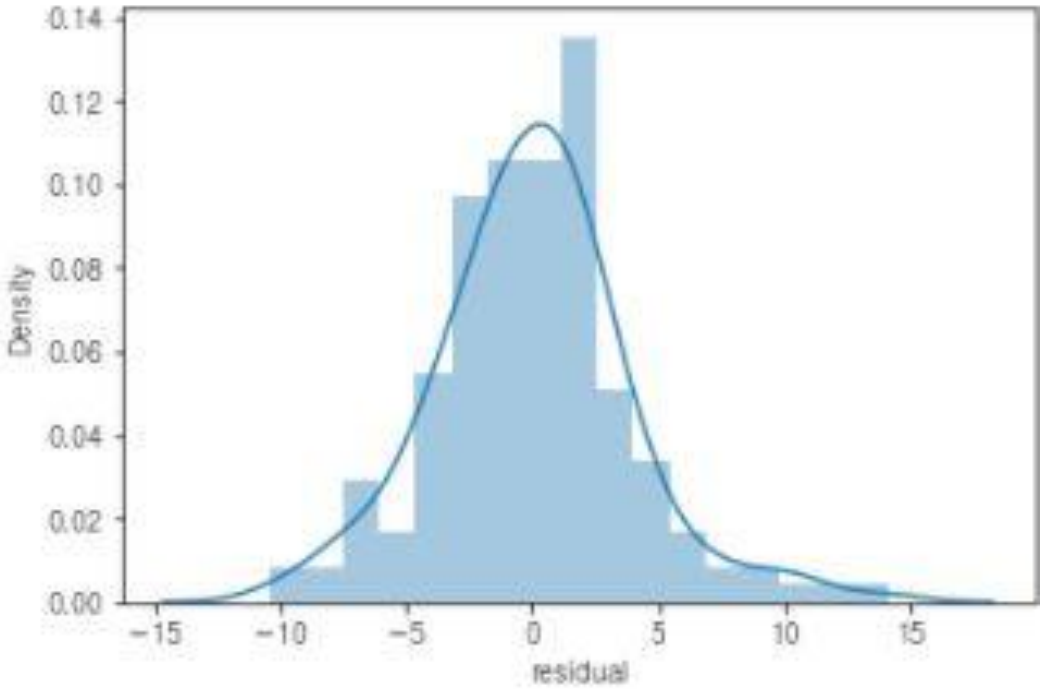
값이 2에 인접할 경우, 오차항의 상관관계가 없는 것

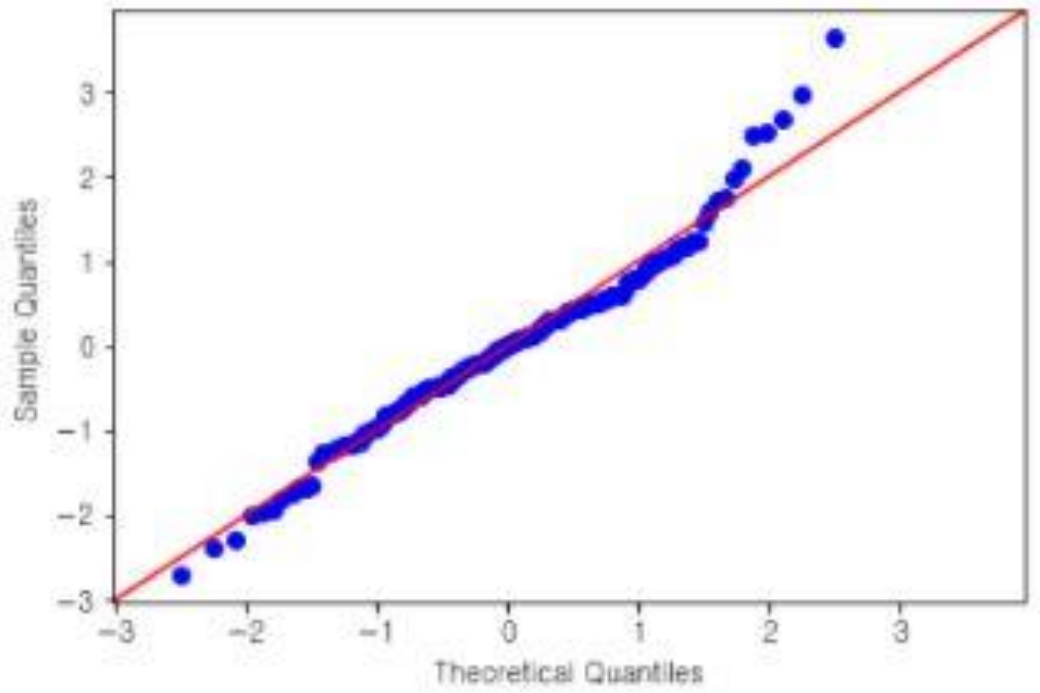
독립성 만족!

3. 모형설계 - 다중선형회귀를 이용한 등락률 예측

▷ 선형회귀분석의 잔차 3가지가정

3. 오차항의 정규성

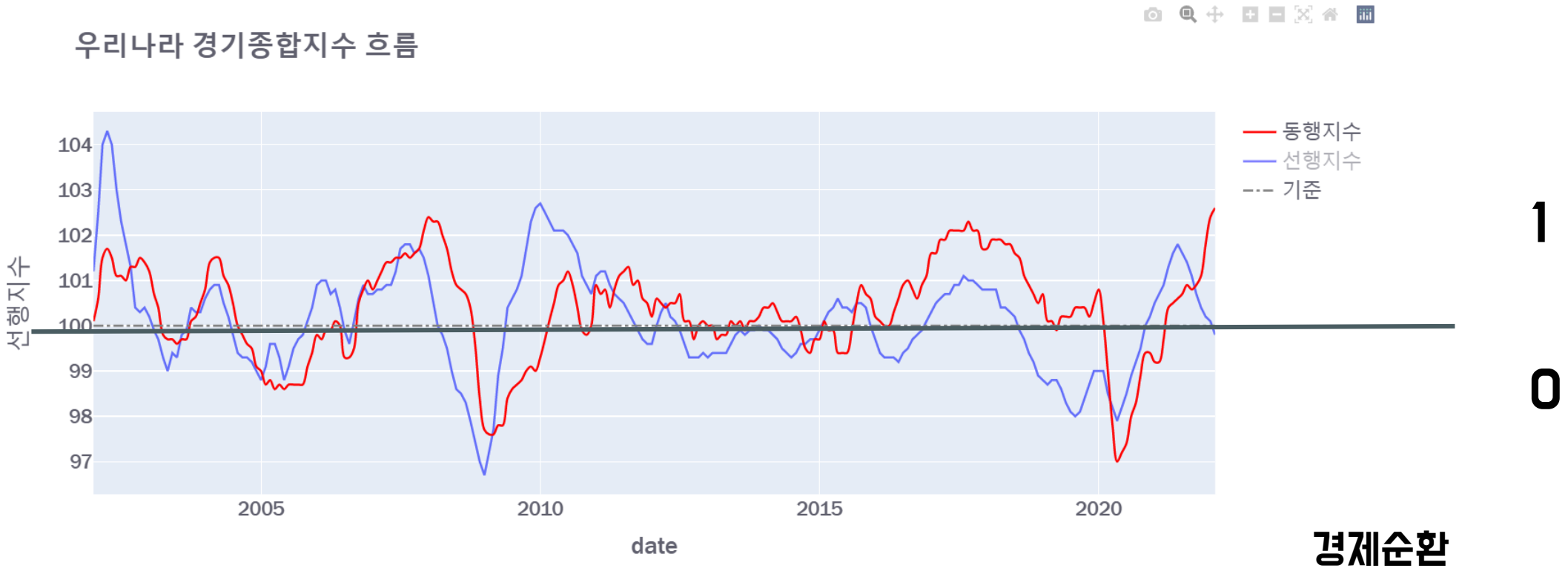




잔차가 완벽히 정규성을 따른다고 하기에는 무리가 있음.
잔차 정규성은 ▲

경기국면별 분석

선행지수 동행지수 시각화



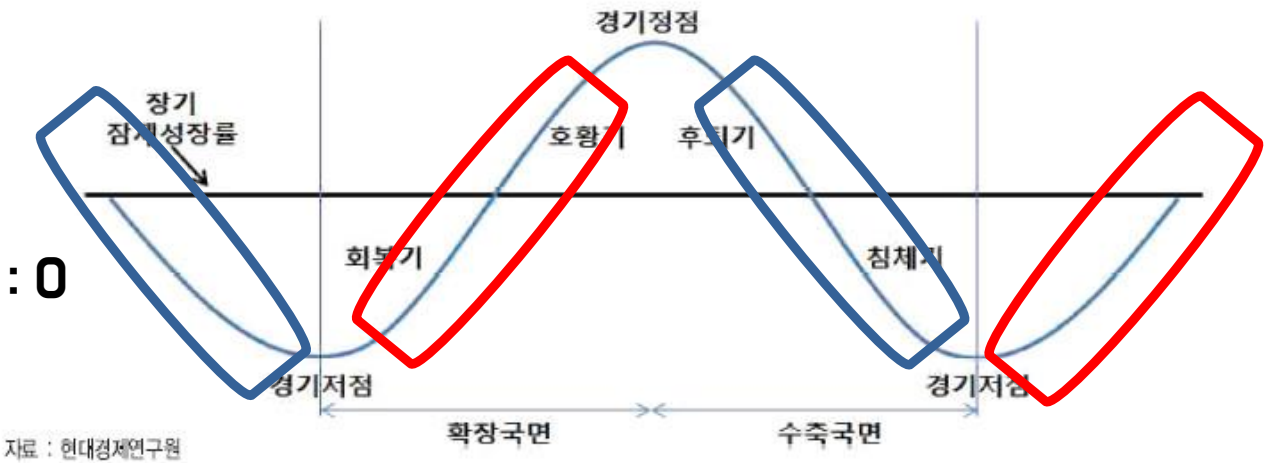
1

0

경제순환

등락률 양수 : 1

등락률 음수 : 0



3. 모형설계 - 경기국면별 등락률 예측

▷ 최종 데이터 셋 다시 보기 | 경기국면별로 모형설계

월별

다중선형회귀 종속변수

날짜	년도	월	전월대비 등락률	등락률(%)	등락률label	외국인	기관	개인	시가총액	동행지수	100기준동행지수	동행지수등락률	동행등락label
2005-01-31	2005	1	0.041053	4.105277	1	8538	984	-9523	42.9	99.0	0	-0.100000	0
2005-02-28	2005	2	0.084336	8.433577	1	14654	-8928	-5725	42.7	98.7	0	-0.303030	0
2005-03-31	2005	3	-0.045167	-4.516690	0	-20741	16920	3820	42.2	98.8	0	0.101317	1
2005-04-29	2005	4	-0.056313	-5.631266	0	-3243	955	2291	41.9	98.6	0	-0.202429	0
2005-05-31	2005	5	0.064644	6.464395	1	1048	17254	-18304	41.7	98.7	0	0.101420	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2021-11-30	2021	11	-0.044323	-4.432316	0	11342	-27380	16641	33.0	101.1	1	0.198216	1
2021-12-30	2021	12	0.048834	4.883389	1	15250	15846	-30452	33.5	101.8	1	0.692384	1
2022-01-28	2022	1	-0.105556	-10.555634	0	-389	6687	12674	32.7	102.4	1	0.589391	1
2022-02-28	2022	2	0.013457	1.345673	1	-6207	-3629	-10135	32.4	102.6	1	0.195312	1
2022-03-31	2022	3	0.021662	2.166212	1	-23373	6022	-3604	31.6	102.4	1	-0.194932	0

206 rows × 13 columns

가변수처리 | 로지스틱회귀 종속변수

가변수 처리 | 독립변수

▷ 선형회귀를 이용한 등락률 예측

경기종합지수 분류 – 동행지수 100이상

OLS Regression Results

Dep. Variable:	등락률(%)	R-squared:	0.353
Model:	OLS	Adj. R-squared:	0.341
Method:	Least Squares	F-statistic:	29.47
Date:	Tue, 03 May 2022	Prob (F-statistic):	6.13e-11
Time:	23:26:24	Log-Likelihood:	-300.73
No. Observations:	111	AIC:	607.5
Df Residuals:	108	BIC:	615.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2640	0.356	0.741	0.460	-0.442	0.970
외국인순매수	9.301e-05	2.04e-05	4.566	0.000	5.26e-05	0.000
개인순매수	-0.0001	3.27e-05	-3.357	0.001	-0.000	-4.49e-05

Omnibus:	11.769	Durbin-Watson:	2.006
Prob(Omnibus):	0.003	Jarque-Bera (JB):	25.236
Skew:	0.324	Prob(JB):	3.31e-06
Kurtosis:	5.244	Cond. No.	2.10e+04

경기동행지수 1000이상 데이터

```
df_u=df[df['100기준동행지수']==1]
df_u
```

R-Squared : 0.353

F검정, P검정 → 모든 변수 유의 0

$$y = 9.301 * e^{-5} * 외국인 - 0.0001 * 개인 + 0.2640$$

▷ 선형회귀를 이용한 등락률 예측

경기종합지수 분류 – 동행지수 100미만

OLS Regression Results

Dep. Variable:	등락률(%)	R-squared:	0.331
Model:	OLS	Adj. R-squared:	0.304
Method:	Least Squares	F-statistic:	12.38
Date:	Tue, 03 May 2022	Prob (F-statistic):	4.30e-05
Time:	23:16:20	Log-Likelihood:	-155.88
No. Observations:	53	AIC:	317.8
Df Residuals:	50	BIC:	323.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5024	0.653	3.831	0.000	1.191	3.814
외국인순매수	0.0001	3.76e-05	3.793	0.000	6.71e-05	0.000
개인순매수	-0.0001	5.11e-05	-2.469	0.017	-0.000	-2.35e-05

Omnibus:	1.749	Durbin-Watson:	1.939
Prob(Omnibus):	0.417	Jarque-Bera (JB):	1.708
Skew:	0.368	Prob(JB):	0.426
Kurtosis:	2.518	Cond. No.	1.81e+04

경기동행지수 100미만 데이터

```
df_d=df[df['100기준동행지수']==0]
df_d
```

R-Squared : 0.331

F검정, P검정 → 모든 변수 유의 0

$$y = 0.0001 * 외국인 - 0.0001 * 개인 + 2.5024$$

▷ 선형회귀를 이용한 등락률 예측

경기종합지수 분류 – 동행등락label == 1

동행지수 그래프를 그려봤을 때 기울기가 양수

OLS Regression Results

Dep. Variable:	등락률(%)	R-squared:	0.305
Model:	OLS	Adj. R-squared:	0.292
Method:	Least Squares	F-statistic:	22.81
Date:	Tue, 03 May 2022	Prob (F-statistic):	6.10e-09
Time:	23:47:35	Log-Likelihood:	-307.81
No. Observations:	107	AIC:	621.6
Df Residuals:	104	BIC:	629.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8721	0.422	2.065	0.041	0.035	1.710
외국인순매수	0.0001	2.13e-05	4.866	0.000	6.15e-05	0.000
개인순매수	-8.57e-05	3.52e-05	-2.436	0.017	-0.000	-1.59e-05

Omnibus:	9.406	Durbin-Watson:	1.999
Prob(Omnibus):	0.009	Jarque-Bera (JB):	10.254
Skew:	0.549	Prob(JB):	0.00593
Kurtosis:	4.047	Cond. No.	2.23e+04

```
df_uu=df[df['동행등락label']==1]
df_uu
```

✓ 0.2s

R-Squared : 0.305

F검정, P검정 → 모든 변수 유의 0

$$y = 0.0001 * 외국인 - 8.57 * e^{-5} * 개인 + 0.8721$$

▷ 선형회귀를 이용한 등락률 예측

경기종합지수 분류 – 동행등락label == 0

동행지수 그래프를 그려봤을 때 기울기가 음수

OLS Regression Results

Dep. Variable:	등락률(%)	R-squared:	0.273
Model:	OLS	Adj. R-squared:	0.246
Method:	Least Squares	F-statistic:	10.16
Date:	Tue, 03 May 2022	Prob (F-statistic):	0.000180
Time:	23:36:11	Log-Likelihood:	-156.81
No. Observations:	57	AIC:	319.6
Df Residuals:	54	BIC:	325.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2210	0.530	0.417	0.679	-0.842	1.284
외국인순매수	5.981e-05	3.05e-05	1.960	0.055	-1.36e-06	0.000
개인순매수	-0.0001	4.73e-05	-2.720	0.009	-0.000	-3.38e-05

Omnibus:	13.659	Durbin-Watson:	1.646
Prob(Omnibus):	0.001	Jarque-Bera (JB):	15.525
Skew:	0.999	Prob(JB):	0.000425
Kurtosis:	4.596	Cond. No.	2.09e+04

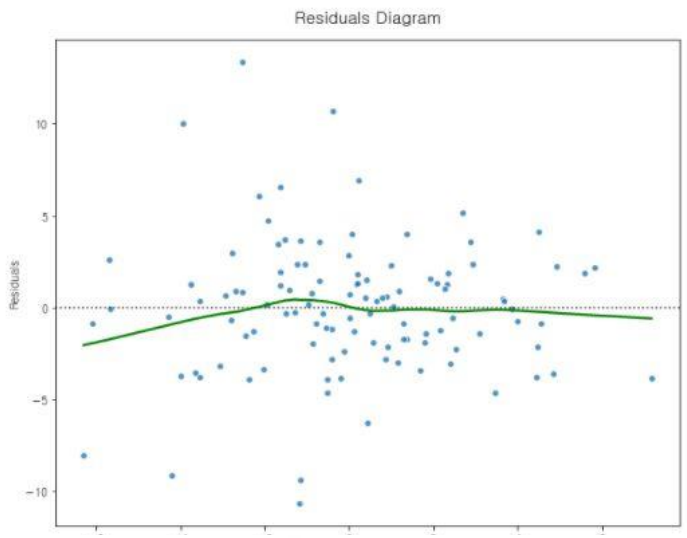
```
df_dd=df[df['동행등락label']==0]
df_dd
```

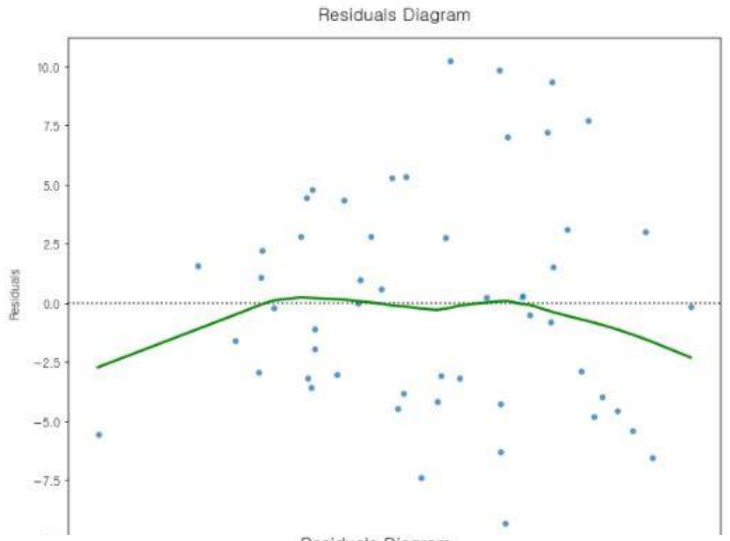
R-Squared : 0.273

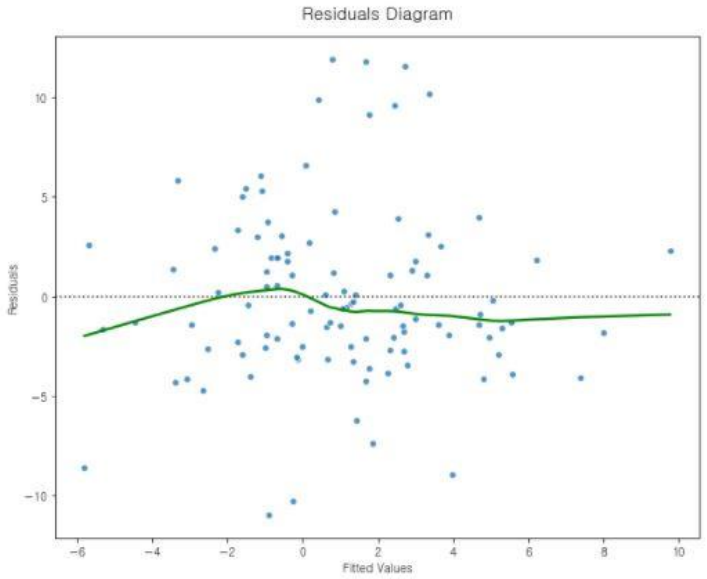
F검정, P검정 → 모든 변수 유의 0

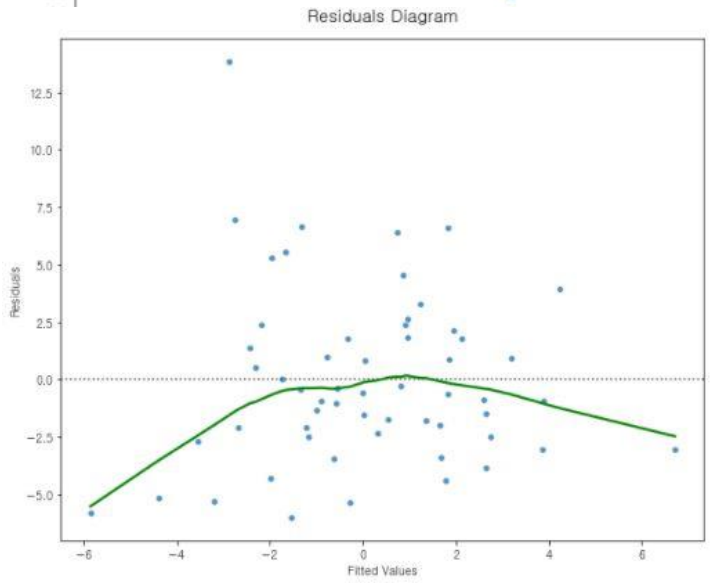
$$y = 5.981 * e^{-5} * 외국인 - 0.0001 * 개인 + 0.2210$$

▷ 오차항의 등분산성



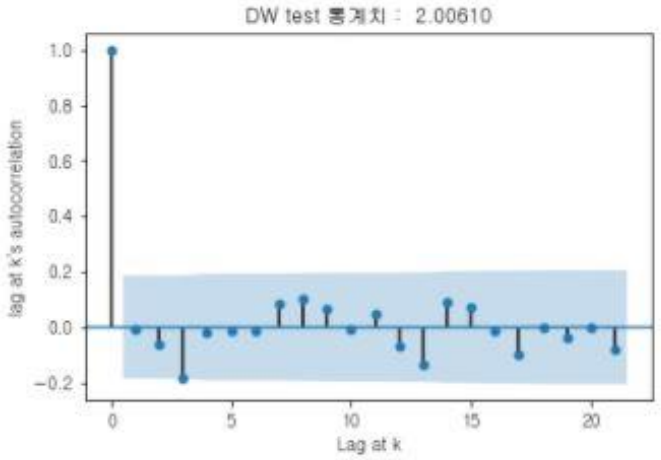


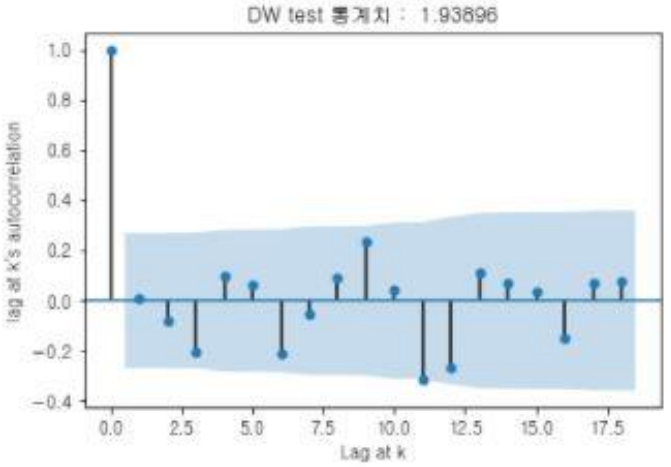


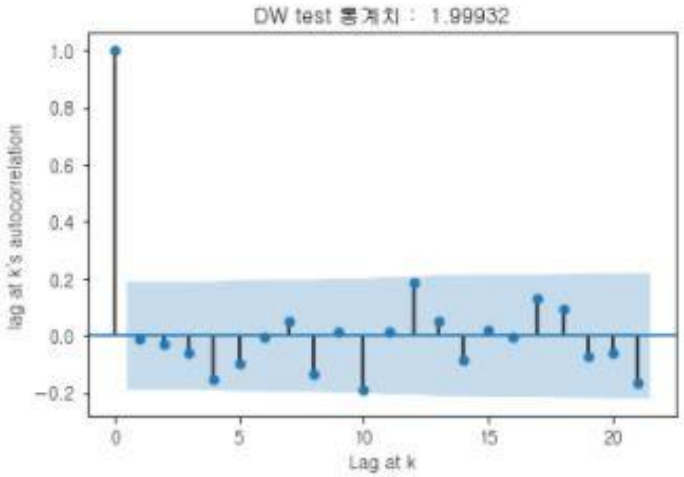


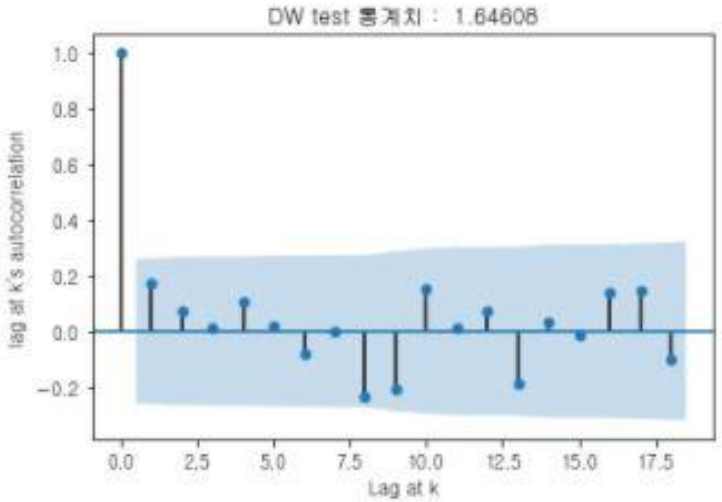
등분산성 O

▷ 오차항의 독립성



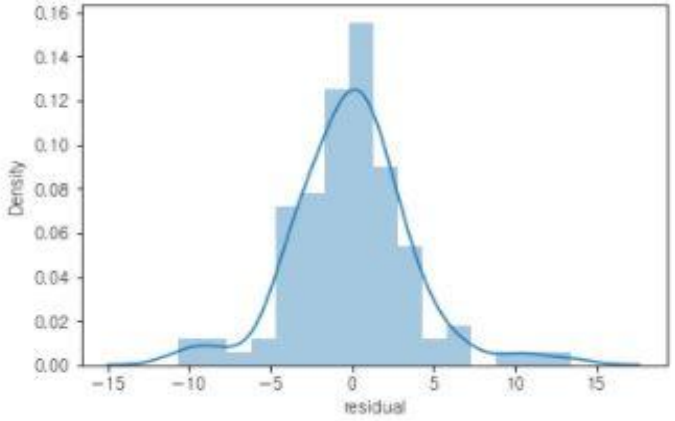


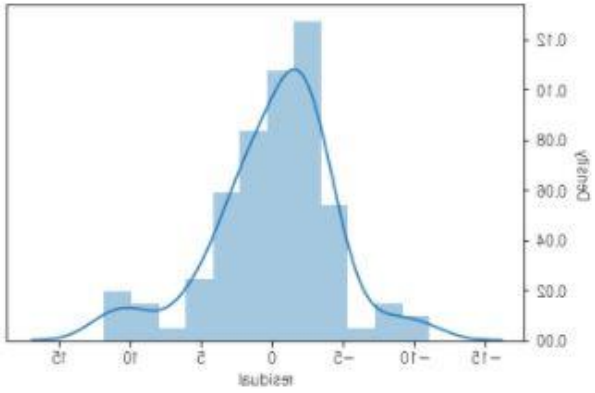


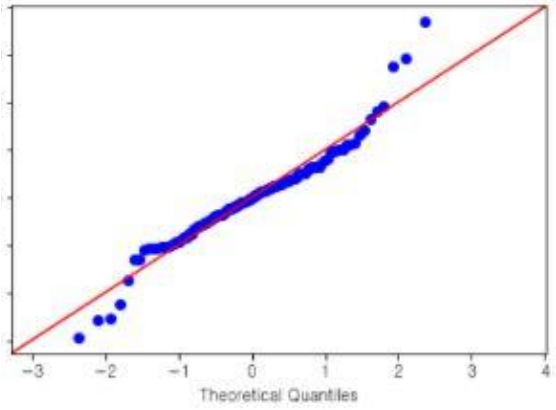


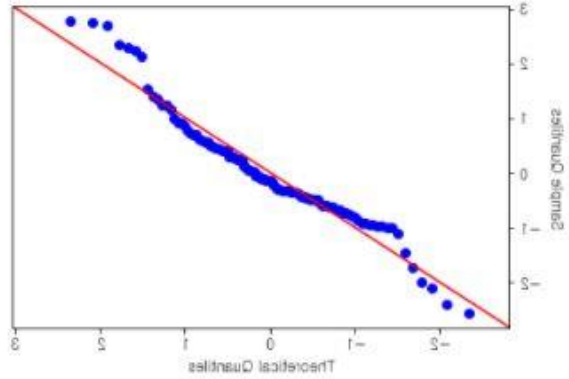
독립성 O

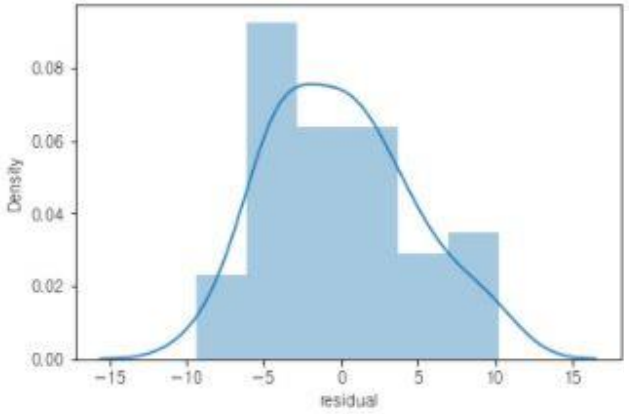
▷ 오차항의 정규성

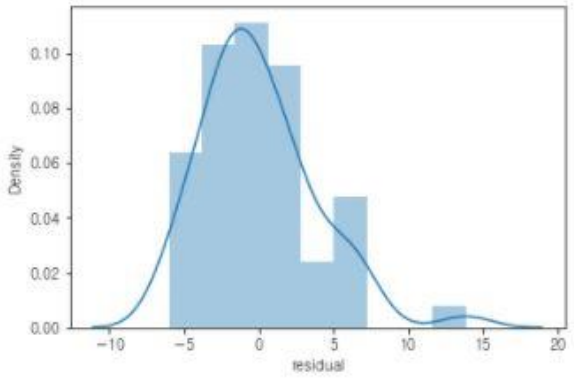

Density
residual

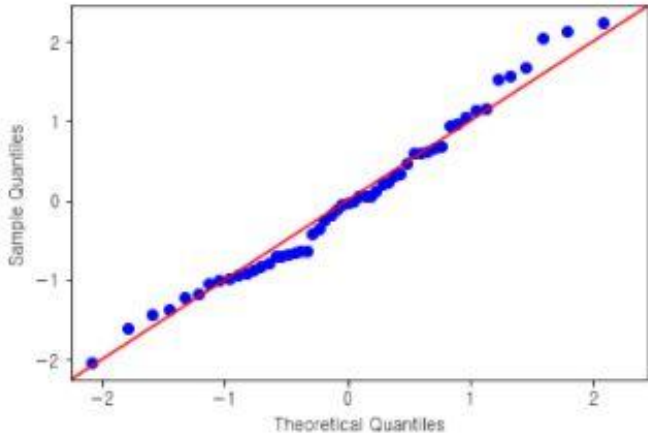

residual

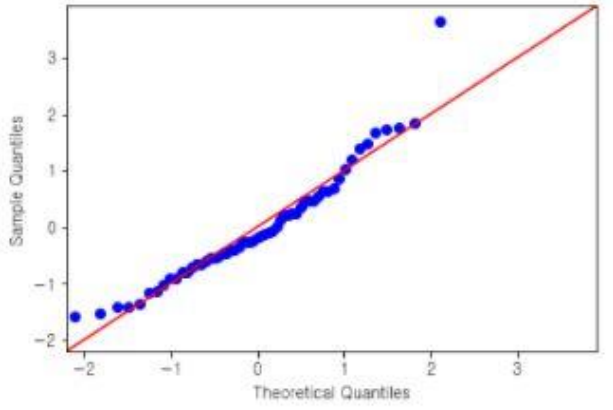

Theoretical Quantiles


Theoretical Quantiles


Density
residual


Density
residual


Sample Quantiles
Theoretical Quantiles


Sample Quantiles
Theoretical Quantiles

▷ 다중공선성 확인

	VIF Factor	features
0	1.172756	외국인순매수
1	1.172756	개인순매수
2	1.005780	const

	VIF Factor	features
0	1.329064	개인순매수
1	1.329064	외국인순매수
2	1.059772	const

	VIF Factor	features
0	1.185044	개인순매수
1	1.185044	외국인순매수
2	1.005953	const

	VIF Factor	features
0	1.274119	개인순매수
1	1.274119	외국인순매수
2	1.041528	const

다중공선성 X

▷ MSE값 확인

```
#MSE 값 구하기
predY2_d = fitted_model_d.predict(X_test_d)
MSE = mean_squared_error(y_true=y_test_d,y_pred=predY2_d)
print(MSE)
```

✓ 0.2s

```
41.36343430393099
```

```
#MSE 값 구하기
predY2_u = fitted_model_u.predict(X_test_u)
MSE = mean_squared_error(y_true=y_test_u,y_pred=predY2_u)
print(MSE)
```

✓ 0.1s

```
10.826036922095623
```

```
#MSE 값 구하기
predY2_uu = fitted_model_uu.predict(X_test_uu)
MSE = mean_squared_error(y_true=y_test_uu,y_pred=predY2_uu)
print(MSE)
```

✓ 0.2s

```
16.279845345969623
```

```
#MSE 값 구하기
predY2_dd = fitted_model_dd.predict(X_test_dd)
MSE = mean_squared_error(y_true=y_test_dd,y_pred=predY2_dd)
print(MSE)
```

✓ 0.1s

```
15.715022614097306
```

선형회귀모형 기본 가정 비교

	정규성	독립성	등분산성	다중 공선성	변수 유의성
전체	O	O	O	X	O
'100기준동행지수' =1(+)	O	O	O	X	O
'100기준동행지수' =0(-)	O	O	O	X	O
'동행등락label'=1 추세상승	O	O	O	X	O
'동행등락label'=0 추세 하락	O	O	O	X	O

3. 모형설계 - 로지스틱회귀를 이용한 등락 예측

▷ 일반화 선형모형 이란?

회귀분석이나 분산분석은 종속변수가 정규분포되어 있는 연속형 변수이다.
하지만 많은 경우에 있어서 종속변수가 정규분포되어 있다는 가정을 할 수 없는 경우도 있으며
범주형 변수가 종속변수인 경우도 있다.

종속변수가 범주형 변수인 경우 : 이항변수(0 또는 1, 합격/불합격, 사망.생존 등)인 경우

→ 종속변수(코스피등락label) 범주형이기 때문에 **Glm을 사용하여 회귀식 도출**

딥러닝을 활용한 로지스틱회귀식 도출code

Generalized Linear Model

```
import statsmodels.formula.api as smf
import statsmodels.api as sm
formula= '등락률label ~ 외국인순매수+개인순매수+기관순매수+시가총액대비+동행지수+동행등락label+동행지수+백기준동행지수'

result2 = smf.glm(formula=formula, data=mtcars, family=sm.families.Binomial()).fit()
print(result2)
print(result2.summary())

glm_pred = result2.predict(mtcars[:5])
print('glm 예측값 :\n', glm_pred)

print('실제값 :\n', mtcars['등락률label'][:5])
glm_pred2 = result2.predict(mtcars)
print('분류 정확도 :', accuracy_score(mtcars['등락률label'], np.around(glm_pred2)))
```

▷ 로지스틱 회귀를 이용한 등락 예측

딥러닝을 활용한 로지스틱회귀식 도출

Generalized Linear Model Regression Results

Dep. Variable:	등락률label	No. Observations:	206
Model:	GLM	Df Residuals:	198
Model Family:	Binomial	Df Model:	7
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-104.79
Date:	Wed, 04 May 2022	Deviance:	209.57
Time:	01:24:27	Pearson chi2:	187.
No. Iterations:	5	Pseudo R-squ. (CS):	0.2894
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.7910	23.463	0.076	0.939	-44.195	47.777
외국인순매수	-5.751e-06	2.36e-05	-0.244	0.807	-5.2e-05	4.05e-05
개인순매수	-0.0001	2.72e-05	-3.683	0.000	-0.000	-4.68e-05
기관순매수	-5.716e-05	2.28e-05	-2.511	0.012	-0.000	-1.25e-05
시가총액대비	0.0543	0.057	0.959	0.338	-0.057	0.165
동행지수	-0.0354	0.237	-0.149	0.881	-0.499	0.429
동행등락label	0.6047	0.357	1.694	0.090	-0.095	1.304
백기준동행지수	-0.2035	0.582	-0.349	0.727	-1.345	0.938

딥러닝을 통해 로지스틱회귀식을 도출

$$\ln\left(\frac{p}{1-p}\right) = -5.751 \times e^{-6} * 외국인 - 0.0001 * 개인 - 5.716 \times e^{-5} * 기관 + 0.0543 * 시가총액대비외국인보유 - 0.0354 * 동행지수 + 0.6047 * 동행등락label - 0.2035 * 100기준동행지수$$

3. 모형설계 - 로지스틱회귀를 이용한 등락 예측

▷ 로지스틱 회귀분석을 통한 KOSPI 등락 예측 [전체 데이터]

로지스틱 회귀분석의 평균 정확도 측정

```
from sklearn.model_selection import GridSearchCV

params = {'penalty': ['l2', 'l1'], 'C': [0.01, 0.1, 1, 1, 5, 10]}

grid_clf = GridSearchCV(lr_clf, param_grid =params, scoring = 'accuracy', cv=3)
grid_clf.fit(data_scaled_ , fd_target)
# StandardScaler()로 평균이 0, 분산 1과 데이터 분포도 현황
scaler = StandardScaler()
data_scaled_ = scaler.fit_transform(_fd)

print('최적 하이퍼 파라미터:{0}, 최적 평균 정확도:{1:.3f}'.format(grid_clf.best_params_, grid_clf.best_score_))
```

```
accuracy: 0.919
roc_auc: 0.914
최적 하이퍼 파라미터:{'C': 5, 'penalty': 'l2'}, 최적 평균 정확도:0.932
```

로지스틱 회귀분석의 분류 정확도 측정

최적 평균 정확도: 0.932

KOSPI 등락 예측 정확도 :93%

3. 모형설계 - 로지스틱회귀를 이용한 등락 예측

▷ 로지스틱 회귀분석을 통한 KOSPI 등락 예측 [전체 데이터]

다른 선형분석 모델과의 성능 지표 비교

```
# model 별로 평가 수행
lrs = LogisticRegression()
lr_reg = LinearRegression()
ridge_reg = Ridge(alpha=10)
lasso_reg = Lasso(alpha=0.01)

for model in [lrs, lr_reg, ridge_reg, lasso_reg] :
  get_model_predict(model, X_train, X_test, y_train, y_test, is_expm1=True)
```

```
### LogisticRegression ###
RMSLE: 0.284, RMSE: 0.488, MAE:0.139
### LinearRegression ###
RMSLE: 0.336, RMSE: 0.652, MAE:0.533
### Ridge ###
RMSLE: 0.332, RMSE: 0.625, MAE:0.525
### Lasso ###
RMSLE: 0.330, RMSE: 0.641, MAE:0.531
```

RMSLE / RMSE / MAE 최저값

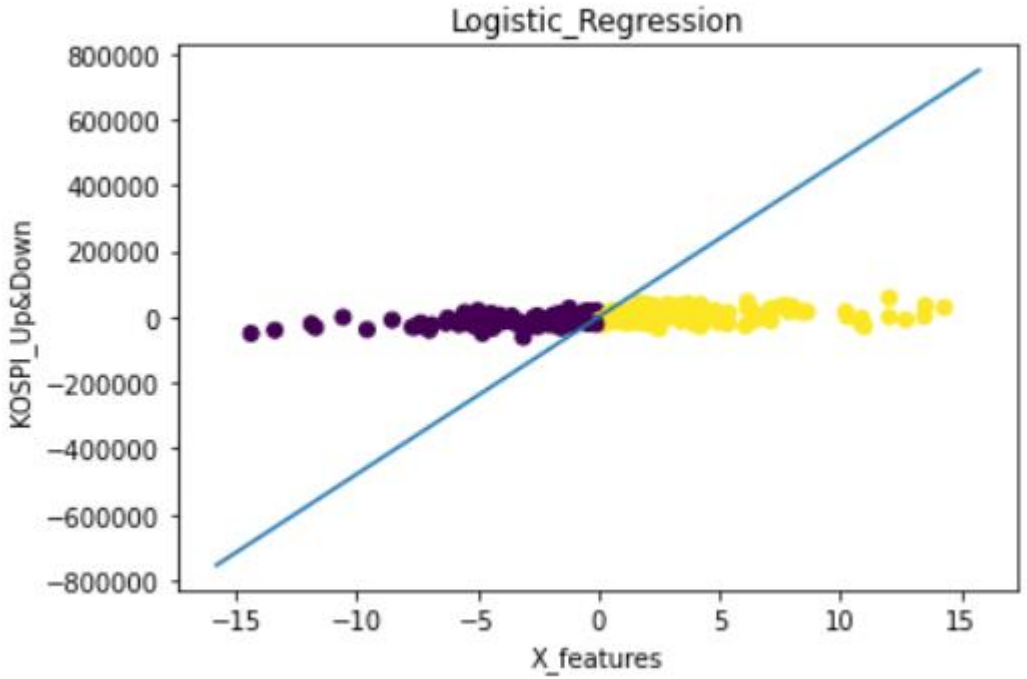
로지스틱 회귀분석의
성능 우수 검증 [전체 데이터]

회귀 분석을 통한 분류 모델이라
등락 예측에 뛰어남

3. 모형설계 - 로지스틱회귀를 이용한 등락 예측

▷ 로지스틱 회귀분석을 통한 KOSPI 등락 예측 [전체 데이터]

산점도로 데이터 분류 시각화



약 93%의
분류 정확도

- 하락 판정 부분은 뛰어난 신뢰도
- 상승 판정 부분은 조금의 아웃라이어 존재

▷ 랜덤포레스트를 통한 feature importance

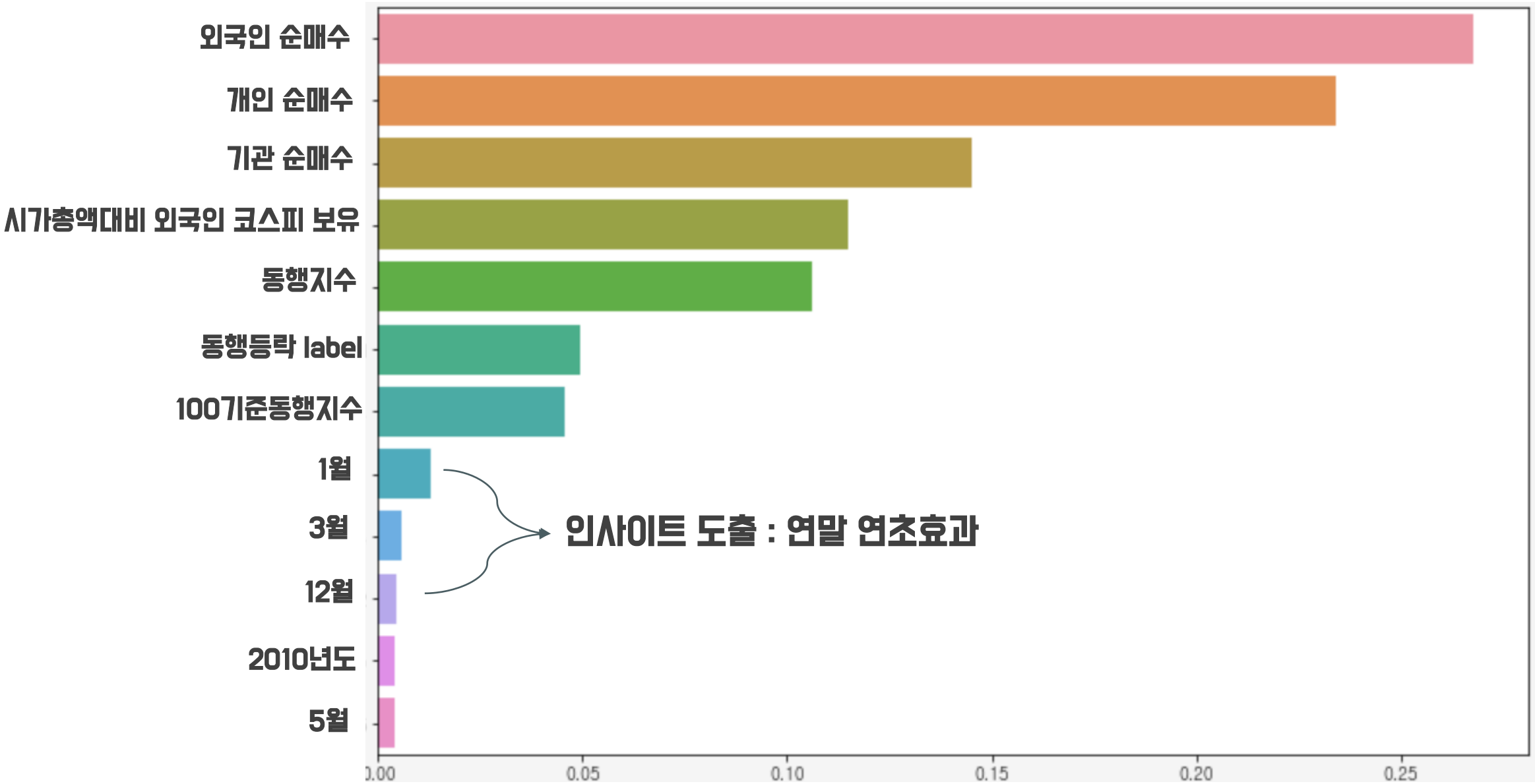
결과를 통해 중요 피쳐 도출

```
1  for_importances_vals=for_ran.feature_importances_
2  for_importances= pd.Series(for_importances_vals,index=X_train.columns)
3  bst_top6=for_importances.sort_values(ascending=False)[:12]
4  bst_top6_featurenames=bst_top6.index
5  bst_top6_featurenames
```

✓ 0.4s

```
Index(['외국인순매수', '개인순매수', '기관순매수', '시가총액대비', '동행지수', '동행등락label', '100기준동행지수',
       '월_1', '월_3', '월_12', '년도_2010', '월_5'],
      dtype='object')
```

3. 모형설계 - 로지스틱회귀를 이용한 등락 예측



3. 모형설계 - 로지스틱회귀를 이용한 등락 예측

▷ 로지스틱 회귀분석을 통한 KOSPI 등락 예측 [features_impotance TOP6 데이터]

로지스틱 회귀분석의 평균 정확도 측정

```
from sklearn.model_selection import GridSearchCV

params = {'penalty': ['l2', 'l1'], 'C': [0.01, 0.1, 1, 1, 5, 10]}

grid_clf = GridSearchCV(lr_clf, param_grid =params, scoring = 'accuracy', cv=3)
grid_clf.fit(df_scaled , df_target)
# StandardScaler()로 평균이 0, 분산 1과 데이터 분포도 현황
scaler = StandardScaler()
df_scaled = scaler.fit_transform(_df)

print('최적 하이퍼 파라미터:{0}, 최적 평균 정확도:{1:.3f}'.format(grid_clf.best_params_, grid_clf.best_score_))
```

```
accuracy: 0.661
roc_auc: 0.657
최적 하이퍼 파라미터:{'C': 1, 'penalty': 'l2'}, 최적 평균 정확도:0.713
```

로지스틱 회귀분석의 분류 정확도 측정

최적 평균 정확도: 0.713

KOSPI 등락 예측 정확도 : 71%

3. 모형설계 - 로지스틱회귀를 이용한 등락 예측

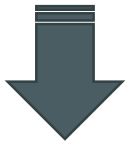
▷ 로지스틱 회귀분석을 통한 KOSPI 등락 예측 [features_impotance TOP6 데이터]

다른 선형분석 모델과의 성능 지표 비교

```
# model 별로 평가 수행
lrs = LogisticRegression()
lr_reg = LinearRegression()
ridge_reg = Ridge(alpha=10)
lasso_reg = Lasso(alpha=0.01)

for model in [lrs, lr_reg, ridge_reg, lasso_reg] :
  get_model_predict(model, X_train, X_test, y_train, y_test, is_expm1=True)
```

```
### LogisticRegression ###
RMSLE: 0.582, RMSE: 1.000, MAE:0.582
### LinearRegression ###
RMSLE: 0.440, RMSE: 0.762, MAE:0.659
### Ridge ###
RMSLE: 0.439, RMSE: 0.761, MAE:0.663
### Lasso ###
RMSLE: 0.442, RMSE: 0.765, MAE:0.668
```

RMSLE / RMSE / MAE 높아짐



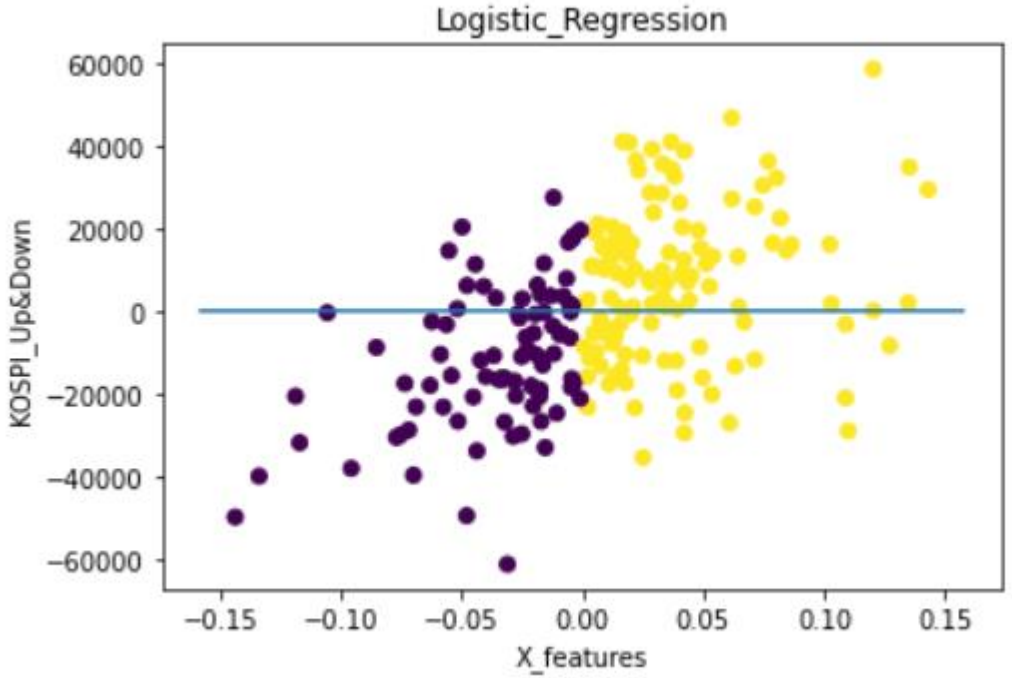
피쳐의 수를 감소시켰더니
정확도가 감소 [TOP6 데이터]

범주형 데이터의 비중이
증가해서 오차가 늘어난 모습

3. 모형설계 - 로지스틱회귀를 이용한 등락 예측

▷ 로지스틱 회귀분석을 통한 KOSPI 등락 예측 [features_impotance TOP6 데이터]

산점도로 데이터 분류 시각화



약 71%의
분류 정확도

- 하락 판정 부분은 아웃라이어가 상대적으로 많다.
- 상승 판정 부분은 상대적으로 아웃라이어가 적다.

결과 해석

4. 결과해석 – 다중선형회귀를 이용한 등락률 예측

▷ 5개 다중회귀모형 비교

전체

$$y = 8.497 * e^{-5} * 외국인 - 0.0001 * 개인 - 0.8017 * 동행지수 + 80.960$$

경기 종합지수 >=100

$$y = 9.301 * e^{-5} * 외국인 - 0.0001 * 개인 + 0.2640$$

경기종합지수 <100

$$y = 0.0001 * 외국인 - 0.0001 * 개인 + 2.5024$$

지수 기울기>0

$$y = 0.0001 * 외국인 - 8.57 * e^{-5} * 개인 + 0.8721$$

지수 기울기 <0

$$y = 5.981 * e^{-5} * 외국인 - 0.0001 * 개인 + 0.2210$$

4. 결과해석 - 다중선형회귀를 이용한 등락률 예측

$$y = 8.497 * e^{-5} * 외국인 - 0.0001 * 개인 - 0.8017 * 동행지수 + 80.960$$

① 회귀직선이 전체 종속변수 값의 변화 중 약 39.5%를 설명함

② 외국인 순매수를 제외한 독립변수들이 고정되어 있을 때,

외국인 순매수량이 1억원 증가할 때 등락률은 0.0008% 만큼 증가한다.

③ 개인 순매수량을 제외한 독립변수가 고정되어 있을 때,

개인 순매수량이 1억원 증가할 때 등락률은 0.0001% 만큼 감소한다.

④ 동행지수를 제외한 독립변수가 고정되어 있을 때,

동행 지수가 1% 증가할 때, 등락률은 0.8017% 만큼 감소한다

4. 결과해석 - 다중선형회귀를 이용한 등락률 예측

경기국면별 전체 비교

	전체	'100기준동행지수' =1(+)	'100기준동행지수' =0(-)	'동행등락label'=1 추세상승	'동행등락label'=0 추세 하락
행 수(데이터수)	206 rows	139 rows	67 rows	134 rows	72 rows
R-Squared	0.395	0.353	0.331	0.301	0.273
MSE	21.67912	10.82603	41.36343	16.27984	15.71502

독립변수로 동행지수를 포함한 전체 기간 모델이 매매주체에 따른 등락률 모델에 가장 정확함.

동행지수가 100을 넘을때 , 동행지수가 추세상승 일때 모형이 더 정확함.

4~5번 추세 모델보다 2~3번 경기 상승/하강 모델이 더 정확하다고 볼 수 있으며,
이는 추세보다우리나라 경기 자체가 상승, 하락일 때 더 코스피 등락률에 영향이 있다는 결론을 도출 할 수 있다

한계점 및 아쉬운 점

로지스틱회귀모형과 머신러닝모형은 적합시켜봤지만

등락을 완전히 "예측" 하는 모형을 만드는 것은 무리

변수간의 다중공선성을 포함한 여러 가정들을 전부 만족하는 변수만 추출하려다 보니
우리가 가진 변수로 실제로 딥러닝 예측모델 -> 정확도 40%

시간이 더 있었다면 .....

- ✓ 코스피에 영향을 미치는 물가지수, 소비자지수 등 여러 경제 지표들을 활용해 예측 모형을 만들어 봤을 것.
- ✓ "등락률"과 "등락 " 을 예측한 모형을 만들었기 때문에 이 둘을 활용한 앙상블 모형 -> 날짜 , 년도, 월 데이터도 있기 때문에 특정 월에서 코스피를 예측하기

전처리로 시작해

전처리로 끝난다

THANK YOU