DFM을 통한 주택가격지수 분석 및 예측

2023020369 장희중

Ⅰ. 분석목적

한국은행의 최근 보고서 '우리나라의 가계부채와 소득불평등'에 따르면 우리나라의 가계대출이 주택 관련에 집중되어 소비 진작의 효과는 미미하고 장기적으로 소득불평등을 확대시킨다고 한다. 이론적으로 가계의 차입은 소비 진작 효과가 큰 것으로 평가되나 우리나라의 대출 목적은 주택으로 대표되는 비금융자산의 취득에 집중되어 오히려 소비 효과를 낮추게 된다. 또한 부채를 통한 자산, 즉 주택의 획득이 장기적으로 고소득층의 미래 소득을 확대하기 때문에 장기적으로 소득불평등을 심화시키게 된다. 따라서 주택가격의 변동요인을 파악하고 이를 통한 정부의 합리적인 정책 마련이 사회적 분위기에 중요한 역할을 할 것으로 보이며 주택매매자 또한 이러한 요인들을 파악하는 것이 중요할 것이다.

주택매매가격의 변동 요인을 파악하는 것에 더하여 동향을 예측하는 것 또한 중요하다. 이는 주택매매자의 매매시기 및 대출에 도움을 줄 것이며 정부의 정책 방향을 결정하는 데에도 기여할 수 있다.

Ⅱ. 수집 데이터

데이터의 수집 기간은 2013년 1월 부터 2023년 9월 까지이며 ECOS와 KOSIS에서 주택매매가격지수에 영향을 미칠만한 변수들을 수집하였다. 데이터 개수는 총 27개이다.

1) ECOS

주택매매가격지수 (전국 총 지수), 고용률 (경제활동인구), 주택담보대출금리 (예금은행 대출금리), 건설업_업황전망, 건설업_매출전망, 건설업_자금사전전망, 소비자 물가지수, 생산자 물가지수, 미분양 주택 현황, M2 (통화량), 국고채(3년, 10년, 30년), 회사채(3년), KOSPI 종가, KOSDAQ 종가, 국채 거래량, 회사채 거래량, GDP, 주택전세가격지수, 환율

2) KOSIS

주택거래량(아파트)

Ⅲ. 분석 과정

1) Granger Causality 검정을 이용한 변수 선택

수집한 데이터들에 대해서 주택매매가격지수(y) 예측에 변수들(x)의 과거 값이 유의한지 검정하기 위해 Granger Causality 검정을 수행해 보았다. 이 검정의 가설은 다음과 같다.

H_0: x로부터 y로의 인과방향이 존재하지 않는다.

H_1: x로부터 y로의 인과방향이 존재한다고 할 수 있다.

위의 가설로부터 유의수준 10% 하에서 p-value가 0.1보다 작으면 x로부터 y로의 인과방향이 존재한다고 판단하였다. 이에 따라 변수 제거의 대상은 p-value가 0.1보다 큰 변수들이다.

granger causality 검정은 정상성을 만족하는 상태에서 진행해야하므로 변수들에 대한 차분(기준금리, 주택건설인허가실적, 전산업생산지수, 주택담보대출금리, 소비자물가지수, 생산자물가지수, 주택매매가격지수는 2차 차분/ 이외의 변수들은 1차 차분)을 거쳐 정상성을 만족(ADF 검정 이용)하도록 하였다.

주택매매가격지수를 제외한 변수들(x)로부터 주택매매가격지수(y)로의 granger causality 검정 결과, 각각의 p-value는 다음과 같다.

변수명(x)	p-value	변수명(x)	p-value
기준금리	0.0004	국고채(3년)	0.0000
주택건설인허가실적	0.0787	국고채(10년)	0.0066
전산업생산지수	0.0128	국고채(30년)	0.0399
주택담보대출금리	0.0029	회사채(3년)	0.0000
건설업_업황전망	0.0135	KOSPI_종가	0.0011
건설업_매출전망	0.0039	KOSDAQ_종가	0.0059
건설업_자금사전전망	0.0489	국채 거래량	0.5124
소비자물가지수	0.1551	회사채 거래량	0.2821
생산자물가지수	0.0287	GDP	0.006
미분양주택현황	0.0043	주택전세가격지수	0.0000
고용률	0.0000	환율	0.0053
M2	0.0691	주택 거래량	0.0155

따라서 국채거래량과 회사채 거래량은 주택매매가격지수 분석에 유의한 변수가 아니라고 판단하고 제거하였다.

반면에 x와 y 사이의 양방향 인과관계 ($x \leftrightarrows y$)가 존재할 경우 또한, x로부터 y로의 인과방향이 존재한다고 단정짓기 어렵다. 하지만 최종적으로 분석에 사용할 DFM모델은 변수들 간의 내생적 관계를 분석하는 모형이므로 양방향 인과관계가 있는 경우는 변수 제거를 하지 않고 모델에 포함시키기로 결정하였다.

2) Train-test 분리

예측 정확도를 측정하기 위해 데이터를 train set과 test set으로 분리하였다. train set은 2013년 3월부터 2022년 8월까지이고 test set은 2022년 9월부터 2023년 9월까지이다.

3) Minmax Scaling

변수들의 척도가 모두 다르기 때문에 MinMax Scaler를 사용하여 값을 0과 1 사이로 맞추어주었다.

4) Principal Component Analysis (PCA)

수집한 변수들을 해석의 용이성을 위해 경제적 관점에 따라 주택매매가격지수, 기준금리를 제외하고 총 8가지의 그룹으로 분리한 후, Principal Component Analysis를 이용하여 각 그룹 내에서 차원축소를 진행하였다. 기준금리는 외생변수로 설정하였다.

먼저 변수들의 그룹화는 다음과 같다.

1. 건설업: 주택건설인허가실적, 건설업_업황전망, 건설업_매출전망, 건설업_자금사전전망
2. 부동산 시장: 주택담보대출금리, 미분양주택현황
3. 채권금리: 국고채(3년), 국고채(10년), 국고채(30년), 회사채(3년)
4. 주식시장: KOSPI 종가, KOSDAQ 종가
5. 물가: 소비자물가지수, 생산자물가지수
6. 경제: 환율, 전산업생산지수, 고용률, M2, GDP
7. 주택전세가격지수
8. 주택 거래량

PCA 적용 후, 각 그룹별 설명비율은 다음과 같다.

Explained ratio		
Group	Principal Component	
	PC1	PC2
건설업	0.57	0.26
부동산 시장	0.78	0.22
채권 금리	0.93	0.05
주식 시장	0.94	0.06
물가	0.86	0.14
경제	0.45	0.18

6개 그룹 모두, 해석의 간결성을 위해 첫번째 principal component만 사용하기로 결정하였다.
건설업과 경제 그룹은 다소 낮은 설명력을 가지고 있어 좀 더 합리적인 데이터 수집이 필요한 점이 연구의 보완점으로 꼽힌다.

PCA 후, 재정의된 데이터이다. 이 데이터를 이후 DFM분석에 적용시킨다.

Date	건설업_pc1	부동산 시장_pc1	채권금리_pc1	주식시장_pc1	물가_pc1	경제_pc1	주택전세가격지수	주택 거래량	주택매매가격지수
2013-01-01	-0.741866	0.778400	0.613032	-0.440554	-0.235296	-0.730172	0.000000	0.000000	0.004551
2013-02-01	-0.451541	0.742935	0.576023	-0.368113	-0.200063	-0.710044	0.009709	0.068153	0.003208
2013-03-01	-0.523430	0.692743	0.484398	-0.351124	-0.216168	-0.553820	0.024272	0.175387	0.002161
2013-04-01	-0.368974	0.665479	0.448577	-0.356738	-0.231702	-0.436456	0.043689	0.281171	0.002226
2013-05-01	-0.422549	0.601515	0.481964	-0.321312	-0.242049	-0.410529	0.053398	0.374761	0.001899

5) Dynamic Factor Model (DFM)

DFM 모델은 소수의 잠재 요인이 더 많은 수의 관측된 시계열의 공통적인 역학을 설명한다고 가정한다. 주택매매가격지수는 여러 복합적 요인에 의해 결정되기 때문에 이 요인들을 추측하여 주택매매가격지수와 함께 DFM모델에 적용시키면 주택매매가격지수와 관련 요인들의 내생적 관계를 잘 분석할 수 있을 것이라 판단하였다.

최종적으로 적용한 DFM 모형식은 다음과 같다.

$$X_t = \Lambda F_t + A x_t + u_t$$
$$F_t = B_1 F_{t-1} + B_2 F_{t-2} + \eta_t$$
$$u_t = C_1 u_{t-1} + C_2 u_{t-2} + \epsilon_t$$

X_t: Endogeonous var (기준금리 제외 모든 변수)
F_t: Latent Factor
x_t: Exogeonous var (기준금리)
u_t: 개별 요인

AIC를 기준으로 factor 개수와 factor lag를 튜닝한 결과, n_factors=1, factor order=2로 설정하였고, 잔차의 자기상관성을 해결하기 위해 error order=2로 설정하였다. 변수들의 정상성을 만족시키기 위해 함수 내에서 'enforce_stationarity=True'로 설정하였다. 이 옵션은 함수 내에서 autoregressive component의 정상성을 만족시키기 위해 AR 파라미터들을 변형시킨다.

아래는 모델 fitting을 위한 코드이다.

```
endog = train_total
exog = X_train_sc[['기준금리']]
n_factors = 1
dfm = DynamicFactor(endog=endog, exog = exog, k_factors=n_factors,
factor_order=2, enforce_stationarity=True, error_order=2)
dfm_results = dfm.fit(maxiter=500)
```

분석 결과, Latent Factor와 lag of Latent Factor의 계수는 다음과 같다.

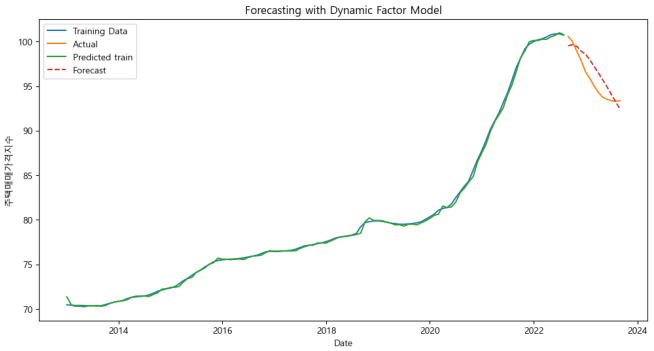
변수	Coef of Latent factor F	P-value
건설업	1.62	0.283
부동산 시장	-2.31	0.0
채권금리	-1.34	0.394
주식시장	2.62	0.126
물가	2.13	0.0
경제	2.9	0.011
주택 전세가격지수	0.23	0.08
주택 거래량	0.50	0.69
주택매매가격지수	0.16	0.031

Lag of Latent factor	Coef of lag.F	P-value
F_{t-1}	0.95	0.0
F_{t-2}	0.04	0.24

유의수준 10% 하에서, 유의한 변수들은 부동산 시장, 물가, 경제, 주택전세가격지수, 주택매매가격지수이며 이 변수들이 유의한 내생적 관계를 가지고 있다고 할 수 있다.

Lag of Latent Factor는 lag1만 유의하게 나왔다.

학습 후, train data와 test data의 예측값에 대한 그래프이다. 예측정확도는 RMSE를 기준으로 하였으며 Train RMSE는 0.2145, Test RMSE는 1.46이다.



6) 모델 비교

전통적인 시계열모델 ARIMA와 딥러닝 모델인 LSTM과 성능을 비교해보았다.

먼저 ARIMA는 auto-arima를 사용하여 fitting 한 결과 order of p,d,q는 (0,2,0) 이며 test RMSE는 4.369로 나왔다. ARIMA에 비해 DFM의 RMSE가 크게 작은 것으로 보아 DFM의 예측 성능이 좋다고 할 수 있다.

두 번째로 딥러닝 모델인 multi input LSTM을 학습시켜보았다. epochs=50, batch_size=1로 설정한 후, fitting 하여 예측한 결과 test RMSE가 0.15로 DFM의 RMSE보다 작게 나왔다.

단순히 지수의 예측이 목적이라면 DFM보다 LSTM을 쓰는 것이 더 좋은 선택이지만 본 분석은 주택매매가격지수에 영향을 끼치는 요인을 분석하는 것이 주요 목적이었기 때문에 DFM을 쓰는 것이 더 합리적이다.

Test RMSE		
DFM	**ARIMA**	**LSTM**
1.46	4.369	0.15

Ⅳ. 결론 및 한계점

1) 결론

주택매매가격은 건설업, 부동산시장, 물가 등 여러 복합적 요인으로 결정되며 DFM 모형은 이 복합적 요인들의 효과를 계수추정치들을 통해 잘 보여준다. 또한 정확한 집값이 아닌 주택매매가격지수의 예측이기 때문에 DFM 모형의 예측정확도는 주택매매가격지수의 미래 동향을 보기에 좋은 편이라 생각되어진다. 다만 아주 정확한 예측을 목적으로 한다면 LSTM과 같은 딥러닝 모형을 사용하는 것이 좋을 것이다.

2) 한계점

더 다양하고 적절한 데이터의 수집으로 좀 더 합리적인 그룹화를 통한 PCA 수행 후 DFM 모형을 적용한다면 주택매매가격의 복합적 요인을 더 구체적이고 정확하게 분석할 수 있을 것이다.