데이터 마이닝

총 인구 예측, 요인분석, KOSPI 지수 시각화

제출일	2021, 04, 15
작성자	정승호

목 차

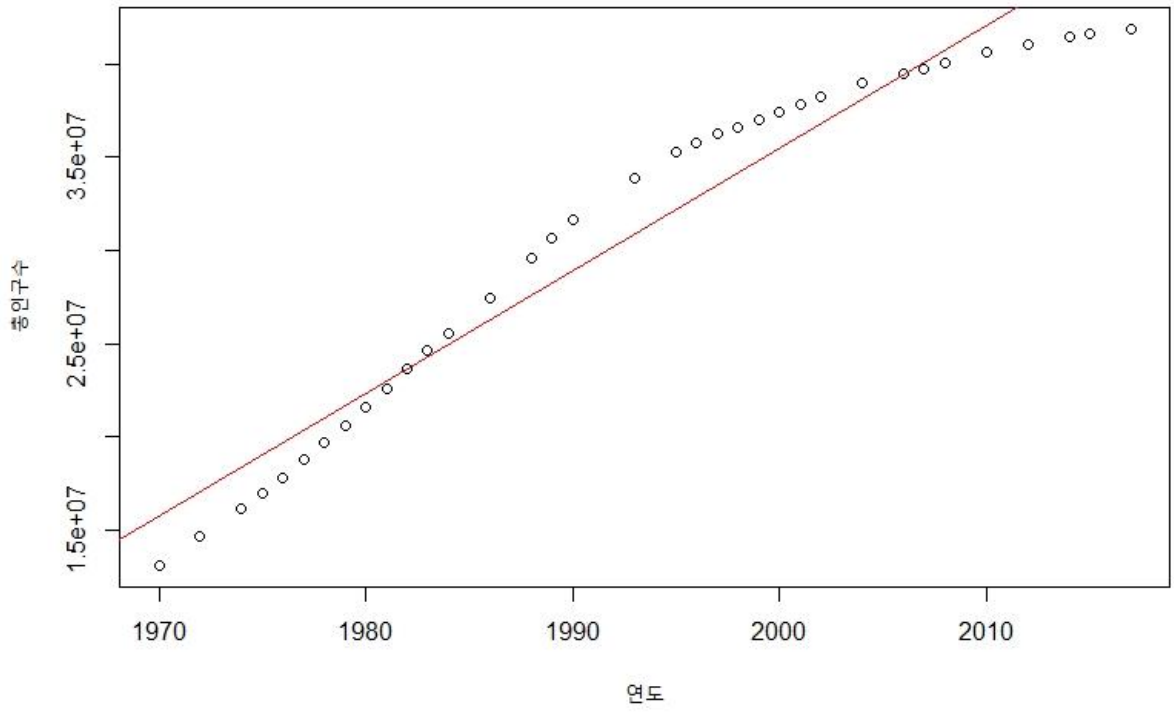
서론

이 레포트는 데이터 마이닝에 관한 내용으로서, 각각 회귀분석과 요인분석, 시계열 분석을 R을 통해 실행한 결과를 다룬다.

본론

1. 대한민국의 2020년과 2021년 총 인구를 예측하시오.

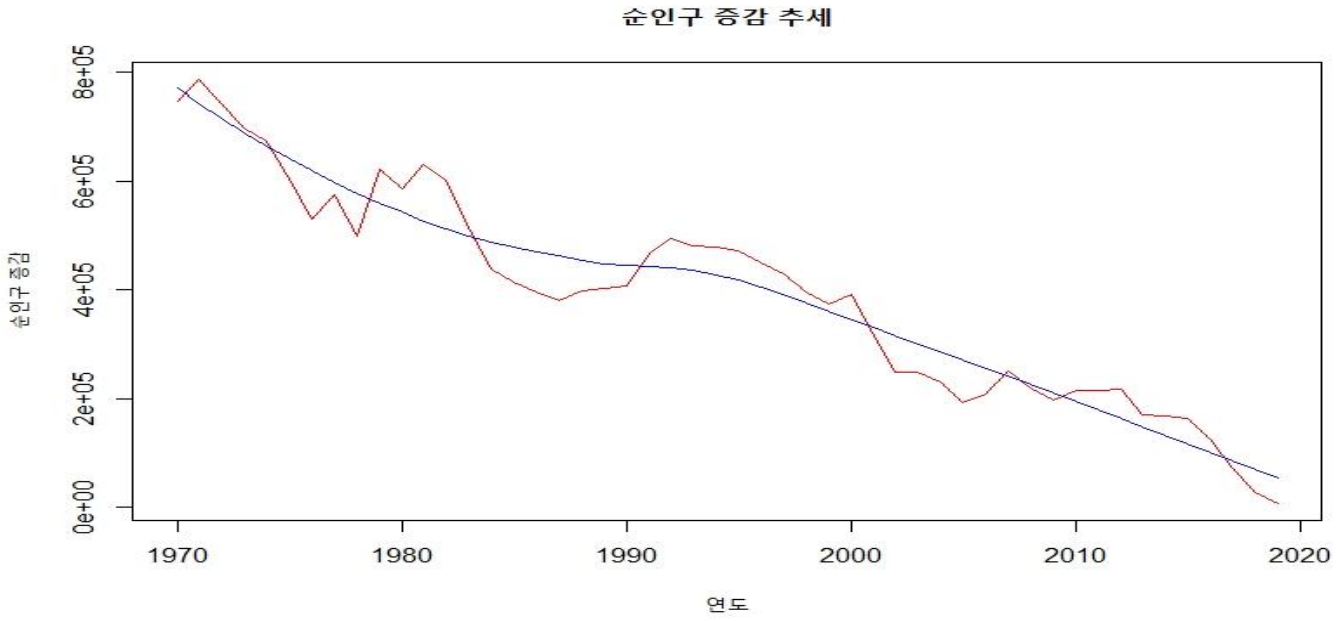
1) 세계은행(https://data.worldbank.org/)에 있는 연도별 우리나라 인구 데이터를 기반으로 회귀분석해 예측.



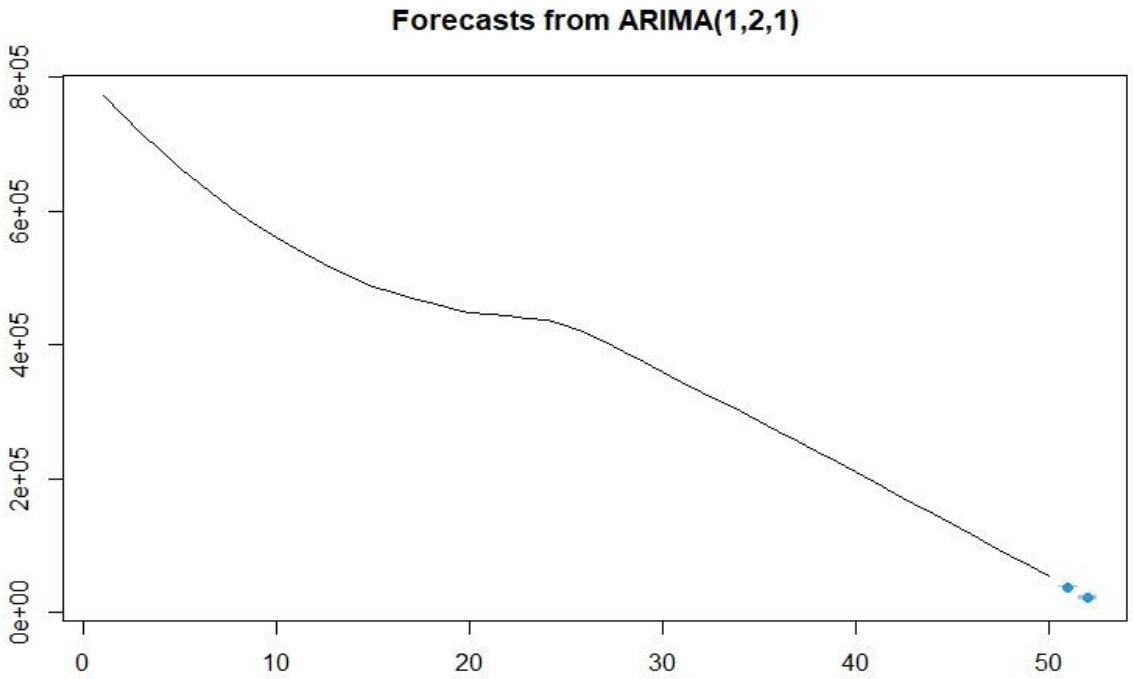
다음과 같은 회귀선을 확인할 수 있었으며 더빈 왓슨 테스트를 통해 잔차에 유의미한 자기상관이 관측되지 않는 것을 확인하였다. ACF(자기상관계수)는 0.97로 매우 높게 나타난다.

회귀분석을 통한 2020년과 2021년 인구수 예측 결과는 각각 **48526000**명**, 49172300**명으로 나타난다.

2) 국가통계포털(https://kosis.kr/index/index.do)에서 과거 연도별 신생아수와 사망자 데이터를 이용하여 인구의 순증감을 계산한 후 과거 연도별 인구의 순증감을 기반으로 시계열분석을 이용하여 예측.



순인구 증감 데이터는 비계절성, 비순환성을 가지고 있으며 시계열 데이터를 시각화한 결과 완연한 하락세를 나타낸다. 추세선을 이용해 시계열 분석을 실시하였으며, 총 인구 증가량과 비교했을 때 잔차의 자기상관이 1건 검출되나 시계열 분석 모형의 적절성을 확인하는 LJung Test 실행결과 0.8이상의 신뢰값이 검출되므로 시계열 분석 모형의 분석결과가 유의미하게 나타났다고 할 수 있다.



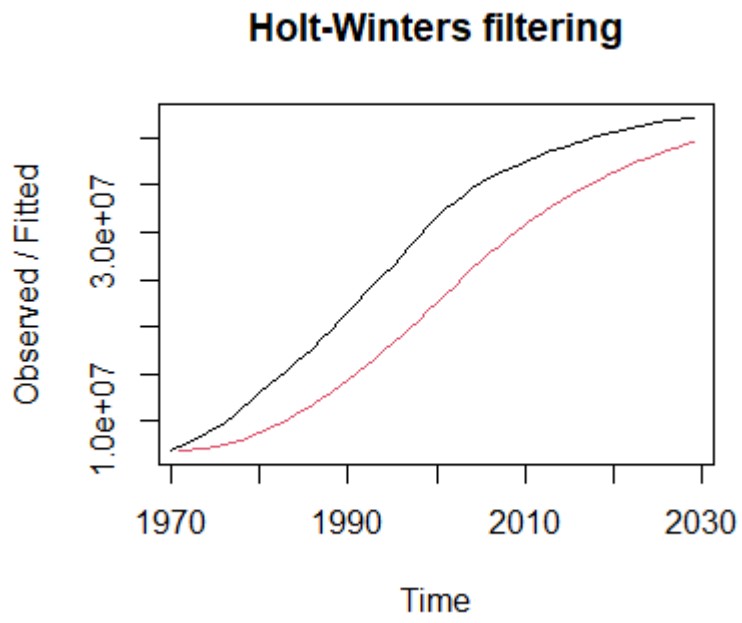
위의 그림처럼 시계열 분석을 통한 대한민국의 총 인구수 예측결과는 꾸준한 하락세로 확인된다.

3) 세계은행에 있는 연도별 우리나라 인구 데이터를 기반으로 Average Annual Growth Rate(AAGR)과 Compound Annual Growth Rate(CAGR)을 계산

Year	AAGR	CAGR			
1970	0.00	0.00	1996	1.50	3.94
1971	5.89	5.89	1997	1.25	3.84
1972	5.34	5.62	1998	1.03	3.73
1973	5.14	5.46	1999	1.02	3.64
1974	4.99	5.34	2000	1.14	3.56
1975	4.87	5.25	2001	1.17	3.48
1976	5.18	5.23	2002	1.03	3.40
1977	5.16	5.22	2003	0.96	3.33
1978	5.00	5.20	2004	0.83	3.25
1979	4.86	5.16	2005	0.64	3.18
1980	4.78	5.12	2006	0.75	3.11
1981	4.59	5.07	2007	0.63	3.04
1982	4.43	5.02	2008	0.89	2.98
1983	4.24	4.96	2009	0.64	2.92
1984	3.88	4.88	2010	0.62	2.86
1985	3.50	4.79	2011	0.76	2.81
1986	3.81	4.73	2012	0.44	2.75
1987	3.84	4.68	2013	0.37	2.70
1988	3.67	4.62	2014	0.54	2.65
1989	3.52	4.56	2015	0.44	2.60
1990	3.36	4.50	2016	0.31	2.55
1991	2.54	4.41	2017	0.21	2.50
1992	2.19	4.30	2018	0.42	2.46
1993	2.12	4.21	2019	0.16	2.41
1994	2.07	4.12			
1995	2.04	4.04			

계산 결과는 각각 위의 결과와 같이 나타난다.

4) 지수평활법 중 홀트윈터스 방식을 이용해 다른 각도로 예측



지수평활법을 사용해 자료를 평활할 경우, 인구수 증가 추세가 완만하게 줄어들지만 감소가 아닌 증가쪽으로 결과가 나타난다. 하지만 홀트윈터스법을 사용해 지수평활한 모델의 평가 상황에서 실제 값과 잔차가 눈에 띄게 큰 정도로 크게 벌어지는 것이 관측되어 사용하기에 부적합하다 판단했다. 지수평활법은 일반적으로 추세가 있는 시계열에 적합하지 않은데, 해당 자료의 경우 눈에 띄는 추세가 관측되기 때문에 사용하기 부적합하다고 판단된다.

2. 아래 표는 SPSS파일인 drinking_water_example.sav파일의 데이터 셋이 구성된 테이블이다. 전체 2개의 요인에 의해서 7개의 변수로 구성되어 있다. 해당 데이터로 요인분석을 실행.

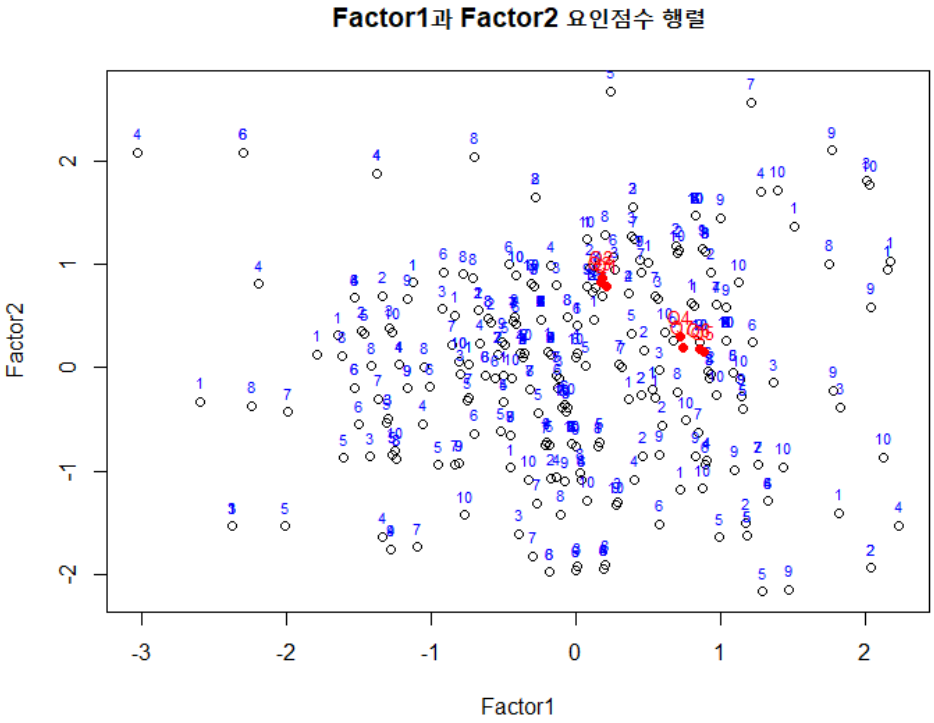
요인구분	변수명(Name)	변수설명(하위요인)	변수값(Value)
제품친밀도	Q1	브랜드	음료의 만족도
	Q2	친근감	
	Q3	익숙함	
제품만족도	Q4	음료의 목 넘김	
	Q5	음료의 맛	
	Q6	음료의 향	
	Q7	음료의 가격	

1) 베리맥스 회전법과 요인점수 회귀분석 방법을 적용하여 요인 분석

```
Call:
factanal(x = drinking.water.df, factors = 2, scores = "regression",     rotation = "varimax")

Uniquenesses:
   Q1    Q2    Q3    Q4    Q5    Q6    Q7
0.333 0.222 0.298 0.388 0.200 0.231 0.410

Loadings:
   Factor1 Factor2
Q1 0.212   0.789
Q2 0.182   0.863
Q3 0.170   0.820
Q4 0.724   0.296
Q5 0.882   0.149
Q6 0.860   0.172
Q7 0.742   0.198

                Factor1 Factor2
SS loadings       2.700   2.219
Proportion Var    0.386   0.317
Cumulative Var    0.386   0.703

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 12.93 on 8 degrees of freedom.
The p-value is 0.114
```

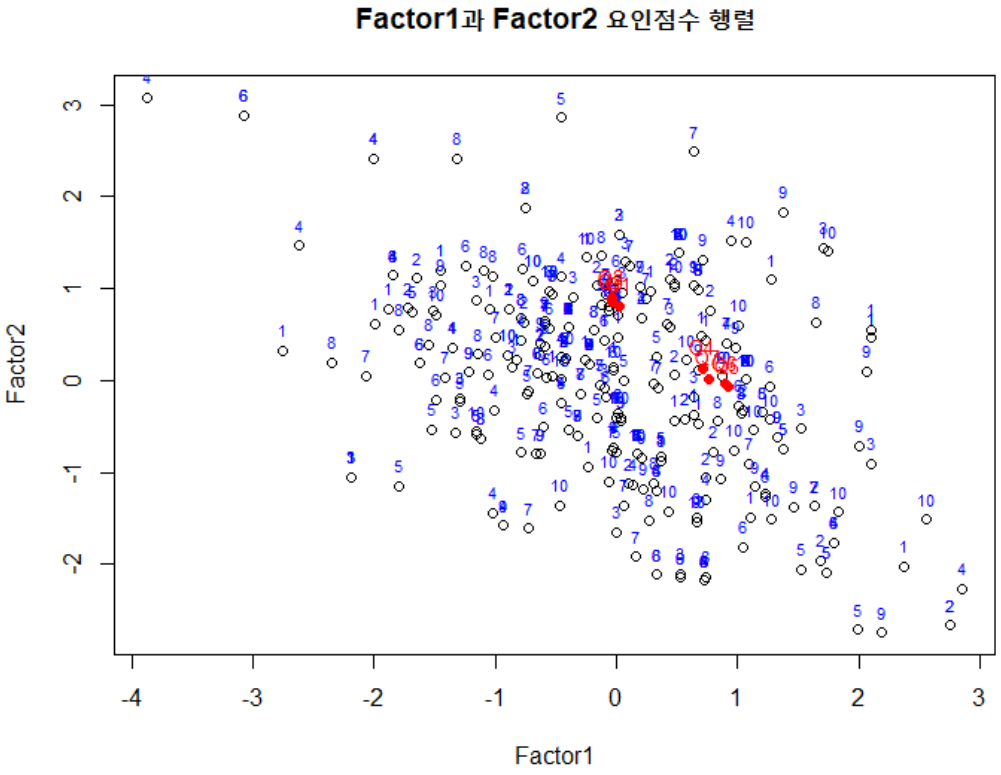
유효성 판단 항목의 값이 통상적으로 0.5 이하이면 유효한 것으로 본다. 그러므로 요인 분석결과 7개의 변수 모두 유효하다고 볼 수 있다. 요인 적재값은 통상 +0.4 이상이면 유의하다고 볼 수 있으며, Factor1에서는 Q4, Q5, Q6, Q7이, Factor2 에서는 Q1, Q2, Q3 변수가 해당된다. 따라서 Q4, Q5, Q6, Q7이 제 1요인으로, Q1, Q2, Q3이 제 2요인이라 할 수 있다. 누적분산비율은 0.703 으로 정보 손실은 약 0.3 이다.

2) 요인점수를 이용한 요인적재량 시각화



위의 분석결과와 같이 2개의 요인으로 묶이는 것을 확인할 수 있다.

3) 프로맥스(Promax)회전법을 적용하여 요인 분석



프로맥스 회전법 역시 2개 요인으로 변수들이 묶이는 결과를 확인할 수 있다.

4) 베리맥스 회전법과 프로맥스 회전법을 적용한 결과 비교 분석

<베리맥스>

```
Uniquenesses:
   Q1    Q2    Q3    Q4    Q5    Q6    Q7
0.333 0.222 0.298 0.388 0.200 0.231 0.410

Loadings:
   제품만족도 제품친밀도
Q1 0.212      0.789
Q2 0.182      0.863
Q3 0.170      0.820
Q4 0.724      0.296
Q5 0.882      0.149
Q6 0.860      0.172
Q7 0.742      0.198

                 제품만족도 제품친밀도
SS loadings        2.700      2.219
Proportion Var     0.386      0.317
Cumulative Var     0.386      0.703

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 12.93 on 8 degrees of freedom.
The p-value is 0.114
```

<프로맥스>

```
Uniquenesses:
   Q1    Q2    Q3    Q4    Q5    Q6    Q7
0.333 0.222 0.298 0.388 0.200 0.231 0.410

Loadings:
   Factor1 Factor2
Q1           0.808
Q2           0.896
Q3           0.853
Q4   0.713   0.132
Q5   0.925
Q6   0.894
Q7   0.759

                 Factor1 Factor2
SS loadings        2.742   2.208
Proportion Var     0.392   0.315
Cumulative Var     0.392   0.707

Factor Correlations:
        Factor1 Factor2
Factor1    1.00    0.46
Factor2    0.46    1.00

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 12.93 on 8 degrees of freedom.
The p-value is 0.114
```

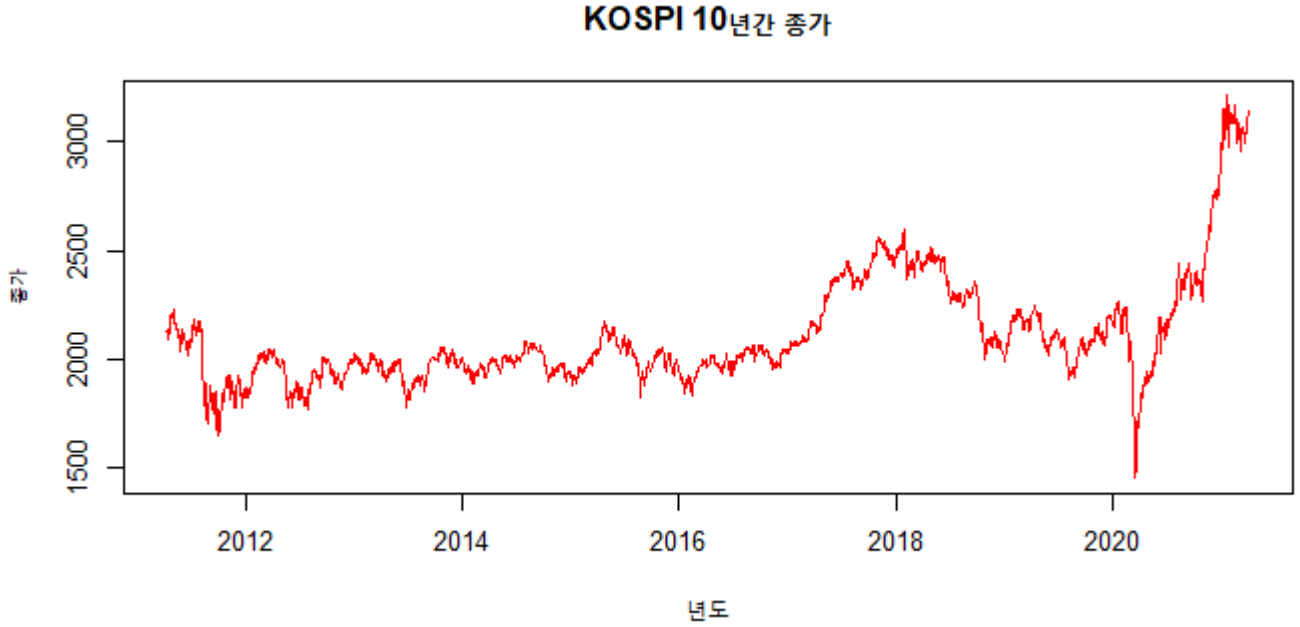
0.5 이하이므로 유효하다

결과 값은 요인적재량(Loadings)를 제외하면 큰 차이가 없는 것을 볼 수 있다. 프로맥스 회전법 사용시 각 Factor에서 다소 높게 측정되었다. 이와 같은 결과는 시각화 결과로 확인해 보았을 때 프로맥스 분석을 했을 경우가 베리맥스 분석보다 변수들이 가깝게 군집한다는 것을 알 수 있다. 또한 프로맥스 분석은 요인간의 상관관계를 나타내 주기 때문에 이를 파악하는데 편리하다.

3. 과거 10년간 일별 KOSPI 지수(종가기준) 데이터를 기준으로 시계열 분석.

(데이터: http://data.krx.co.kr/contents/MDC/MDI/mdiLoader/index.cmd?menuId=MDC0201010101#)

1) 추세선 확인



2) 4가지 시계열 자료의 변동요인을 분해하여 시각화