빅데이터를 활용한 학령인구 예측

권병근[1,*], 성다솜[2], 이다인[2]

[1]부산대학교 수학과, 산업수학 소프트웨어 연계전공, 산업수학센터 학부연구생
[2]부산대학교 수학과, 빅데이터 연계전공, 산업수학센터 학부연구생
이메일 : [*]house9895@pusan.ac.kr

1. Abstract

이 프로젝트는 인구통계학적 데이터와 경제 데이터를 결합한 접근 방식을 사용하여 미래 학령인구를 예측하였다. 통계청(KOSIS)의 연령별 인구 데이터와 국제통화기금(IMF)의 세계 경제 전망(WEO) 데이터를 활용했다. 데이터 전처리에는 수집된 데이터를 분할하고, 학령인구에 필요한 변수를 추가하고, 주요 경제 지표를 선택하는 작업을 진행하였다.

전체 인구 예측을 위한 로지스틱 방정식과 경제 지표 예측을 위한 이차 함수라는 두 가지 모델링을 통하여 미래의 값을 예측하였다. 이후 학령인구를 예측하기 위해 수학적 모델링과 시계열 모델링이라는 두 가지 방법을 진행하였다. 수학적 모델은 이차 다변수 다항함수 함수를 사용하는 반면 시계열 모델은 SAIRMAX(외생변수를 사용한 계절 자동 회귀 통합 이동 평균) 방식을 사용하였다. 결과를 비교하고 시각화하며, 인구의 비현실적인 급격한 감소로 인해 이차 다변수 다항함수를 기각하였고, 우수한 성능과 추세를 보여주는 SARIMAX 모델을 최종 모델로 선정하였다.

최종 모델은 미래 예측값에 대한 독립 변수의 오차 범위를 고려하여 독립변수로는 상대적으로 오차 범위가 적은 총인구수만을 선택하여 종속 변수의 안정성을 확보하였다. 또한 특정 연령 구간의 비율을 구하는 방법을 사용하므로 학령인구뿐만 아니라 영유아 인구, 노년 인구 등 특정 구간의 인구도 예측할 수 있으며, 해당 방법은 전체 학령인구 대비 특정 인구 비율을 예측하는 데에도 적용 가능할 것이다.

2. Introduction

학령인구란 한 나라 또는 지역의 주어진 교육 수준에 이론적으로 대응하는 연령 집단의 인구를 의미하며, 만 6세부터 만 21세까지의 인구를 의미한다. 통계청에서 수집한 연령별 인구 현황 데이터와 IMF에서 수집한 경제 데이터를 전처리한 후, 각각의 데이터를 활용하여 미래를 예측할 수 있는 수리 모델링을 수행하였다. 이후, 수리 모델에서 도출된 예측값을 활용하여 학령인구를 예측하는 작업을 수리 모델링과 시계열 머신 러닝 모델링 두 가지 방법으로 진행하고 최종 모델을 선정하였다.

3. Main

(1) 사용한 데이터

- 연령별 인구 현황 (출처 : 통계청, KOSIS)
- World Economic Outlook data(WEO data) (출처 : 국제금융기구, IMF)

(2) 데이터 전처리

1) 데이터 슬라이싱

수집한 데이터에서 인구 데이터는 1960년부터의 데이터를, 경제 데이터는 1980년부터의 데이터를 가지고 있고 모두 2020년까지만 실측 데이터였다. 따라서 1980년부터 2020년까지의 실제 값인 데이터만 사용하였다.

2) 인구 데이터 – 필요한 데이터 추가

학령인구에 해당하는 연령의 인구의 합인 '학령인구 수' 변수 추가, 학령인구 수를 초 인구수로 나누어 '학령인구의 비율' 변수 추가

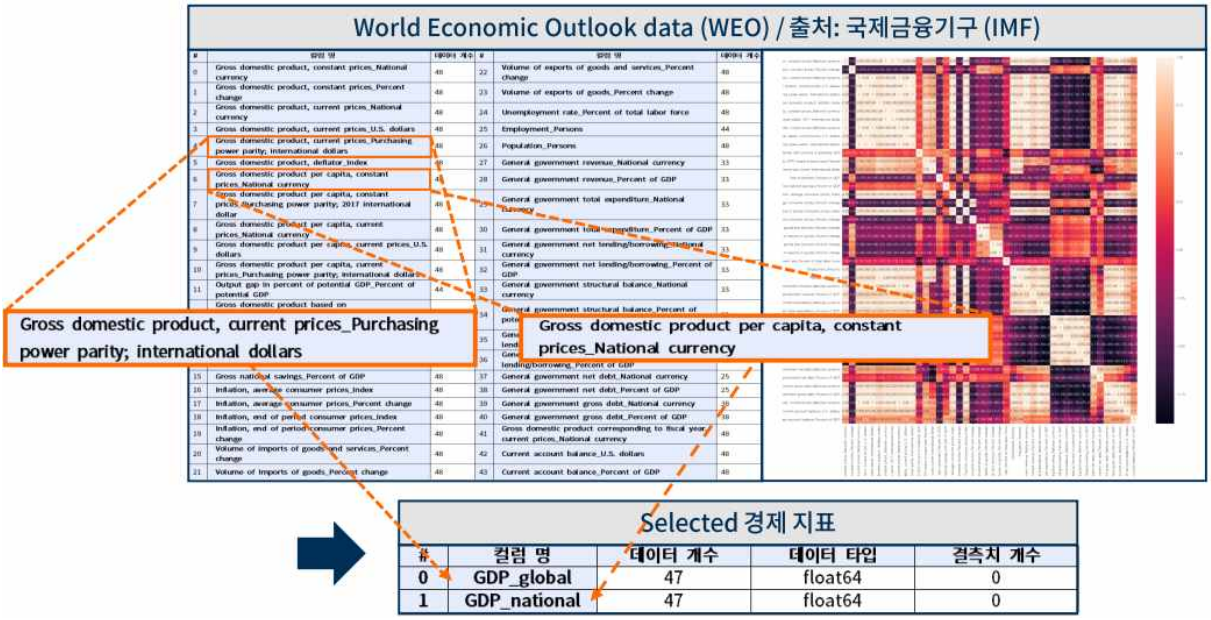
연령별 인구 현황 / 출처 :통계청(KOSIS)				
#	컬럼 명	데이터 개수	데이터 타입	결측치 개수
0	가정별	243	object	0
1	성별	243	object	0
2	연령별	243	object	0
3	항목	243	object	0
4	단위	0	float64	243
5	1960년	243	int64	0
6	1961년	243	int64	0
…	…	…	…	…
114	2069년	243	int64	0
115	2070년	243	int64	0

전체 인구 및 학령 인구				
#	컬럼 명	데이터 개수	데이터 타입	결측치 개수
0	년도	111	int64	0
1	총인구수	111	int64	0
2	학령인구수	111	int64	0
3	학령인구비율	111	float64	0

3) 경제 데이터 – 변수 선택

WEO 데이터에 46개의 feature가 존재하여 상관관계가 높은 feature 들이 다수 존재하였다. 따라서 3가지 기준을 통해 2개의 feature를 선택하였다.

- 결측치가 없는 feature
- 상관관계가 0.9 이상으로 묶인 그룹 중 가장 큰 그룹 2개 선택
- 각 그룹에 속하는 변수 중 다른 feature 들과의 상관계수(소수점 셋째 자리에서 반올림) 가 1인 것이 가장 많은 변수 1개 선택



Gross domestic product, current prices_Purchasing power parity; international dollars와 Gross domestic product per capita, constant prices_National currency 변수 선택 이후에 GDP_global, GDP_national로 feature의 이름을 간략하게 변경하였다.

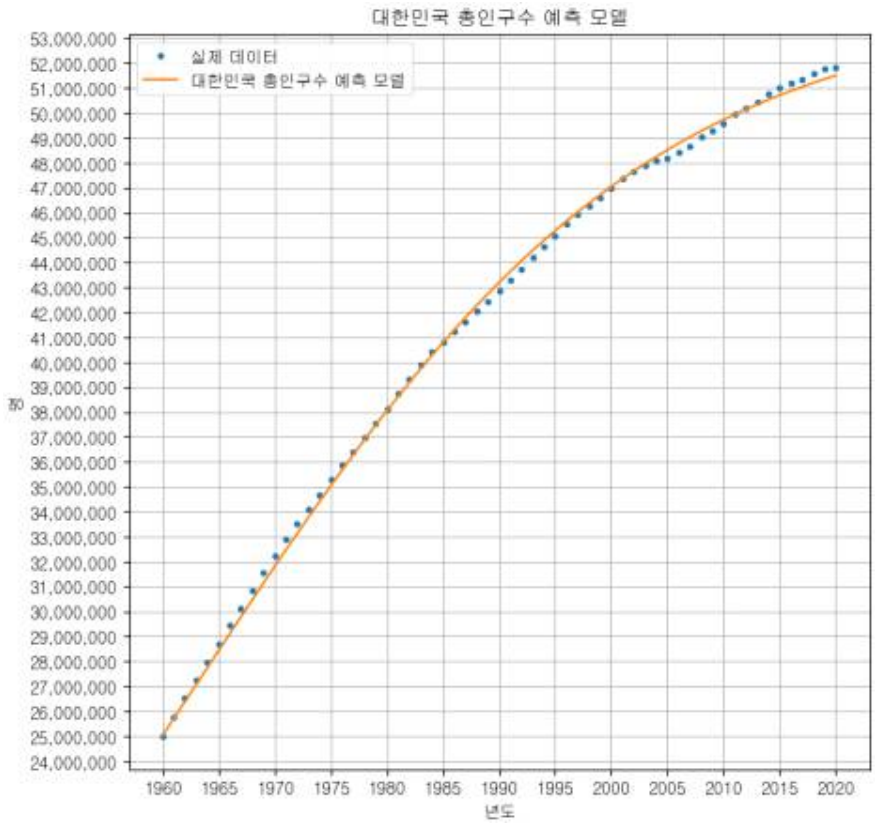
(3) 총인구 예측 수리 모델링

1) Population logistic equation(P.F. Verhulst)

$$\frac{dP}{dt} = P(r - \frac{r}{K}P),\ (P : \text{총인구수},\ r : \text{인구 성장율},\ K : \text{최대 인구})$$

베르누이 인구 로지스틱 방정식으로 총인구를 예측하는 수리 모델링을 하였다. 그리고 실제 데이터를 사용하여 최소 제곱법으로 파라미터를 추정하였고, 아래의 식과 그래프를 얻었다.

$$\frac{dP}{dt} = P(0.050203168979557594 - \frac{0.050203168979557594}{54488908.29241143}P)$$



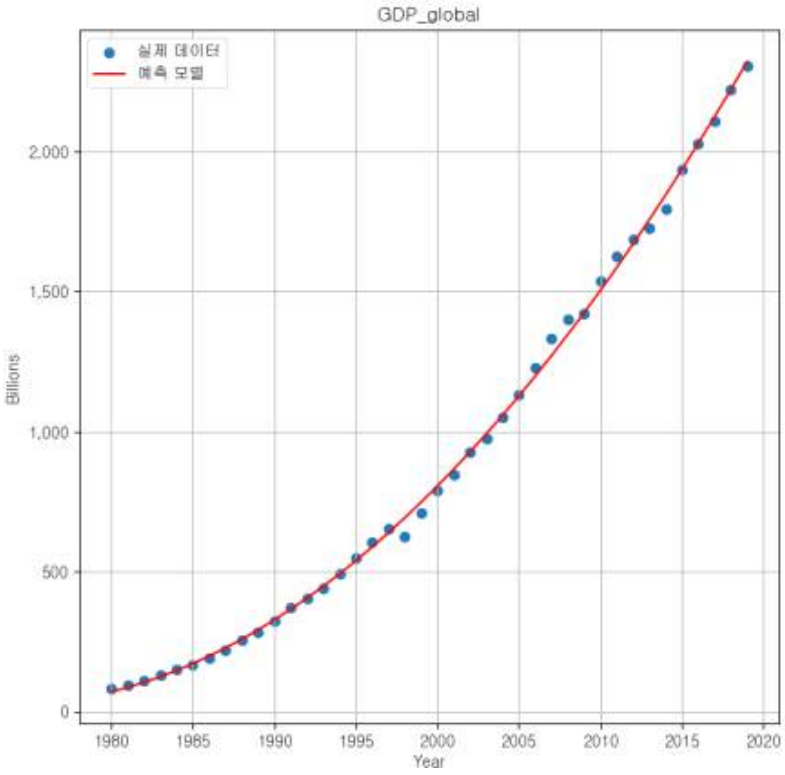
이 모델을 활용하여 2021년부터 2060년까지의 총인구수를 예측하였다.

(3) 경제 지표 예측

1) WEO의 GDP_global 변수 예측

GDP_global을 예측하기 위해 기본적인 $Y = at^2 + bt + c$ 이차 함수식을 모델링 하였다. 그리고 실제 데이터를 사용하여 최소 제곱법으로 파라미터를 추정하였고, 아래의 식과 그래프를 얻었다.

$$Y = 1.1026484419143t^2 - 4351.84852965273t + 4293908.41434967$$
$$(Y:\ GDP\,global,\ t:\ year)$$

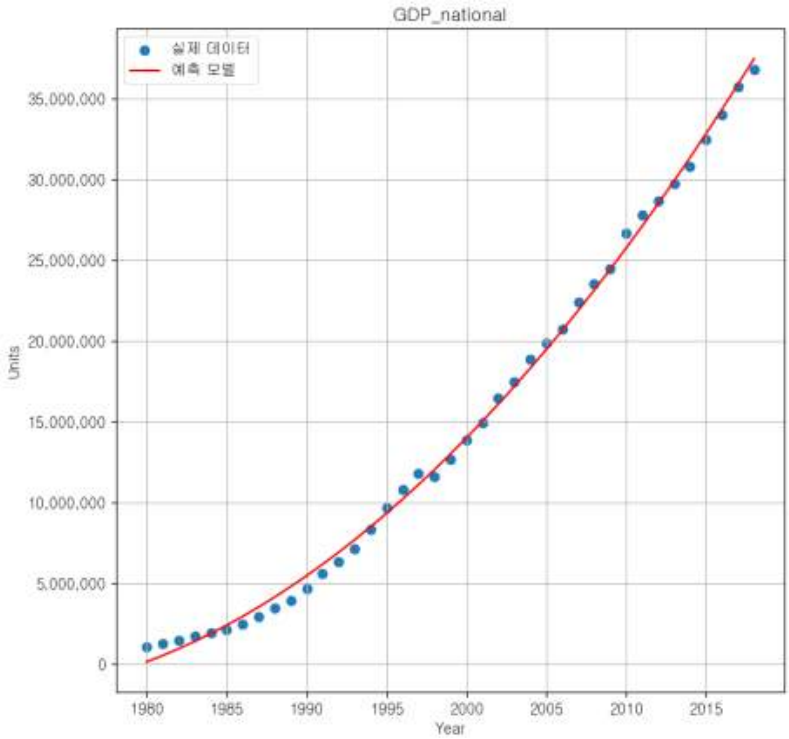


이 모델을 활용하여 2021년부터 2060년까지의 GDP_global feature를 예측하였다.

2) WEO의 GDP_national 변수 예측

GDP_national을 예측하기 위해 기본적인 $Y = at^2 + bt + c$ 이차 함수식을 모델링 하였다. 그리고 실제 데이터를 사용하여 최소 제곱법으로 파라미터를 추정하였고, 아래의 식과 그래프를 얻었다.

$$Y = 16098.1091324323t^2 - 63377457.85787t + 62376475908.611$$
$$(Y : GDP national, t : year)$$



이 모델을 활용하여 2021년부터 2060년까지의 GDP_national feature를 예측하였다.

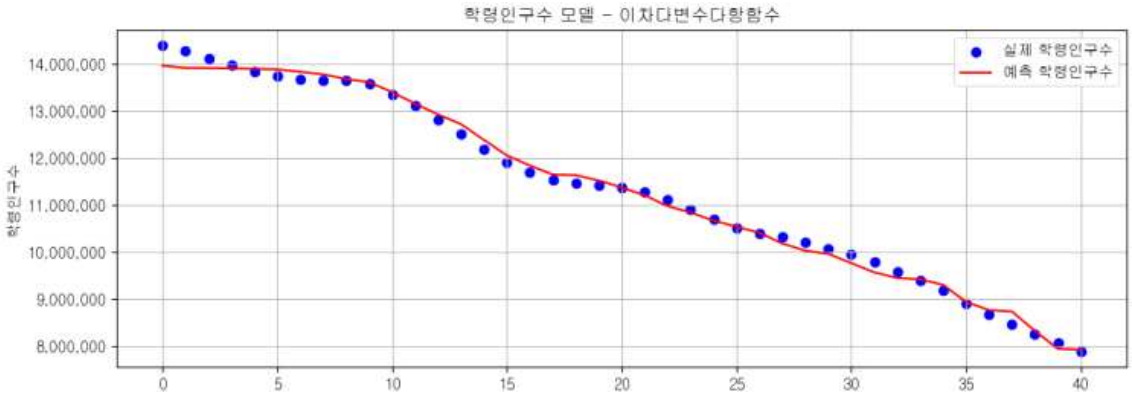
(4) 학령인구 예측

1) 수리 모델링을 이용한 학령인구 예측

학령인구를 예측하기 위해, 앞서 예측한 경제 데이터를 활용하여 이차 다변수 다항함수를 세우고 실제 데이터를 사용하여 최소 제곱법으로 파라미터를 추정하였고, 아래의 식과 그래프를 얻었다.

$$Y = 9.56592274\,x_1^2 + 5.9291.228 \times 10^{-8}\,x_2^2 - 1.54072578 \times 10^{-3}\,x_1 x_2 + 845.23441\,x_1 - 0.706889057\,x_2 + 14009144.2$$

(Y: 학령인구, t : $year$, x_1 : $GDP\,global$, x_2 : $GDP\,national$)



2) 시계열 모델을 이용한 학령인구 예측

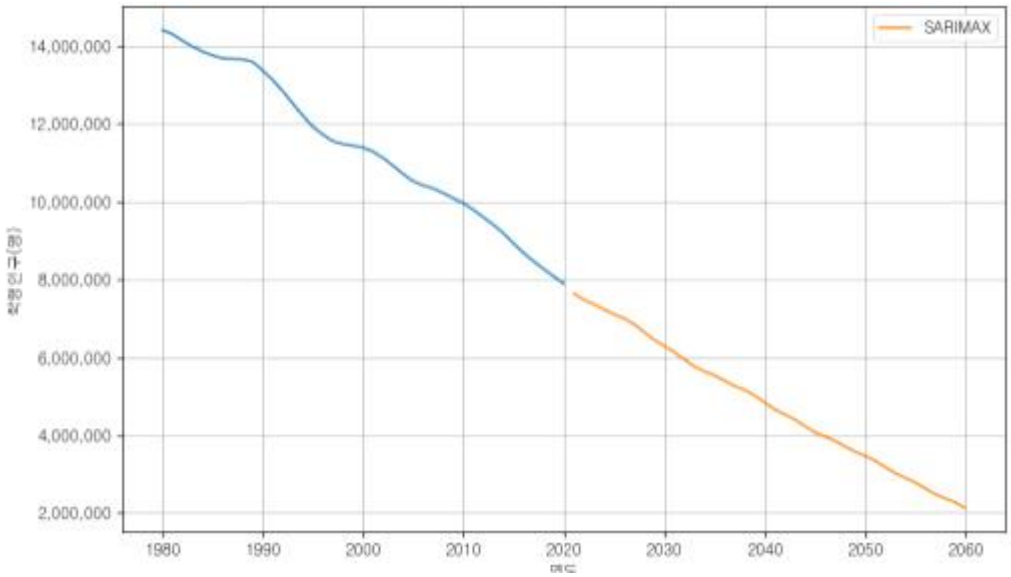
SARIMAX 모델을 사용하여서 학령인구를 예측하였고, 학습에 필요한 파라미터와 외생변수는 아래의 표와 같이 설정하였으며, 모델의 성능 지표인 AIC, BIC는 모두 낮은 값으로 좋은 성능을 보여주었다. 정규 분포, 등분산, 자기 상관성 검증도 진행하여서 모델이 유의미함을 알아내었으며, 생성된 예측값이 추세를 잘 나타내고 있다고 판단하였다.

Non seasonal param	(p, d, q)	(2, 1, 0)
Seasonal parm	(P, D, Q) [m]	(2, 1, 0) [4]
외생변수(exog)	총인구수	
AIC	-47.111	
BIC	-37.610	

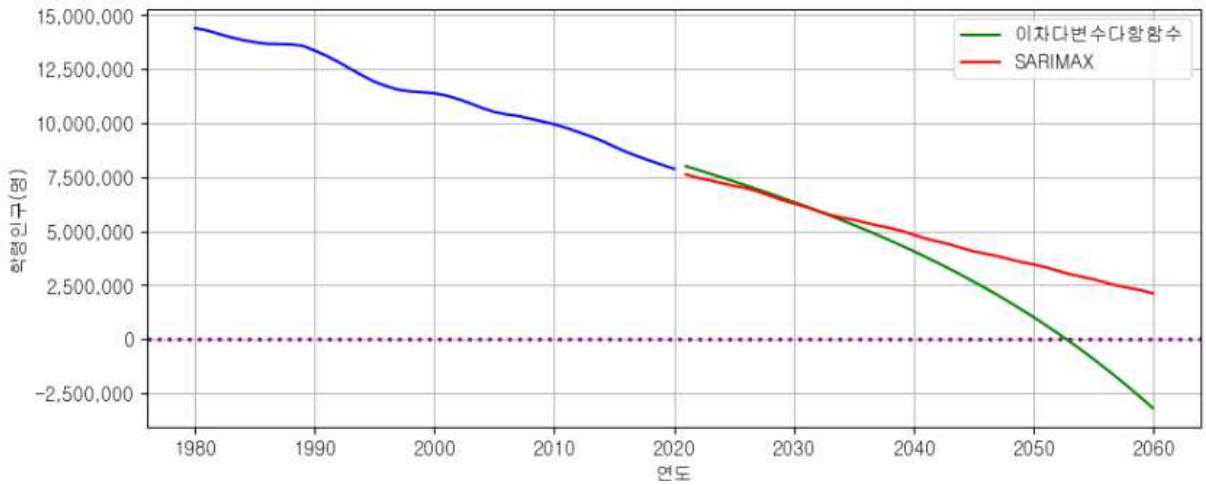
정규 분포 검정	
H0 : 정규 분포 O	H1 : 정규 분포 X
Jarque-Bera의 p-value : 1.00 >=0.05 => H0 채택, 즉 정규 분포이다.	

이분산성 검정	
H0 : 등분산 O	H1 : 등분산 X
Heteroskedasticity의 p-value : 0.90 >= 0.05 => H0 채택, 즉 등분산이다.	

자기 상관성 검정	
H0 : 자기 상관성 X	H1 : 자기 상관성 O
Ljung-Box의 p-value : 0.15 >= 0.05 => H0 기각, 그러나 이보다 더 좋은 값을 찾아내지 못함. 따라서 85% 신뢰 구간에서는 자기 상관성이 있다.	



4. Result & Conclusion



두 가지 방법으로 도출한 예측 모델을 활용하여 1980년부터 2020년의 학령인구와 2021년부터 2060년까지의 예측값을 생성한 뒤 시각화하였다. 이차 다변수 다항함수는 2050년대 초반에 인구가 0 이하로 감소하며 너무 가파르게 감소하여 기각하였다. 따라서 모델 검증 및 수치들이 우수하며 추세도 가장 날 나타냈다고 판단된 SARIMAX를 최종 모델로 선정하였다.

5. Expected Effect

미래 예측값을 위한 독립변수에도 오차 범위가 존재하므로, 독립변수의 증가는 종속 변수의 오차 범위를 확대 시킬 수 있다. 그러나 최종 모델에서는 오차 범위가 상대적으로 적은 총인구수만을 독립변수로 선택함으로써 종속 변수의 안정성을 확보하였다. 그리고 최종 모델은 전체 인구 대비 특정 연령 구간의 비율을 구하는 방법을 사용하므로 학령인구뿐 아니라 영유아 인구, 노년 인구 등 특정 구간의 인구도 예측 가능하다. 또한 해당 방법을 이용하면 전체 학령 인구수 대비 특정 지역의 학령인구 비율을 예측하는 데에도 적용할 수 있다.