2023 K-DS 해커톤
데이터 기반의 학령인구 예측: 수학적 접근

본채만채

권병근 성다솜 이다인

F²m IMC
빅데이터 기반 금융 · 수산 · 제조 혁신 산업수학센터

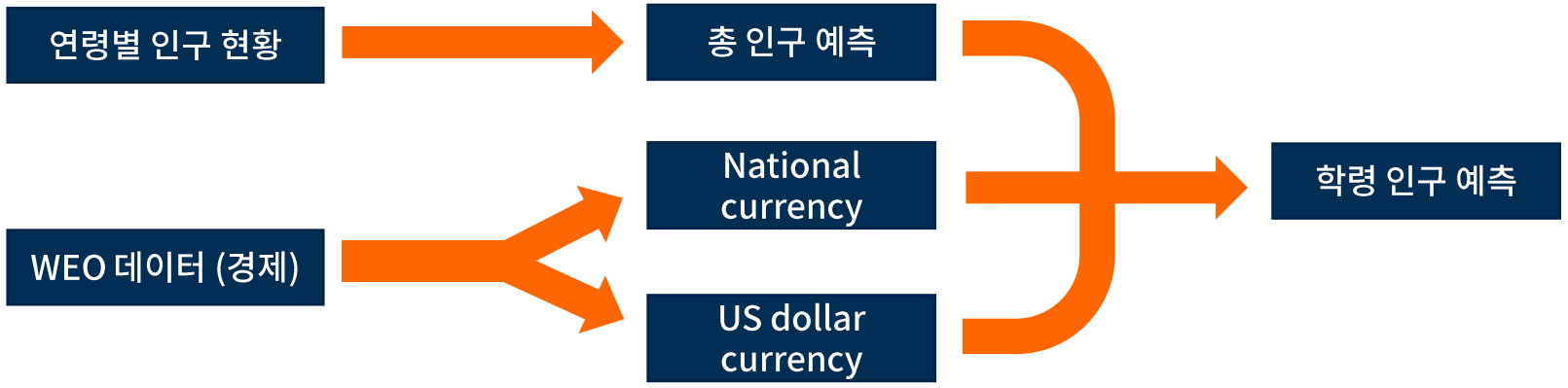
부산대학교 PUSAN NATIONAL UNIVERSITY

Contents

01 아이디어

학령인구란? 한 나라 또는 지역의 주어진 교육 수준에 이론적으로 대응하는 연령 집단의 인구를 의미하며, 만 6세부터 만 21세까지의 인구 (출처: 통계청 KOSIS)

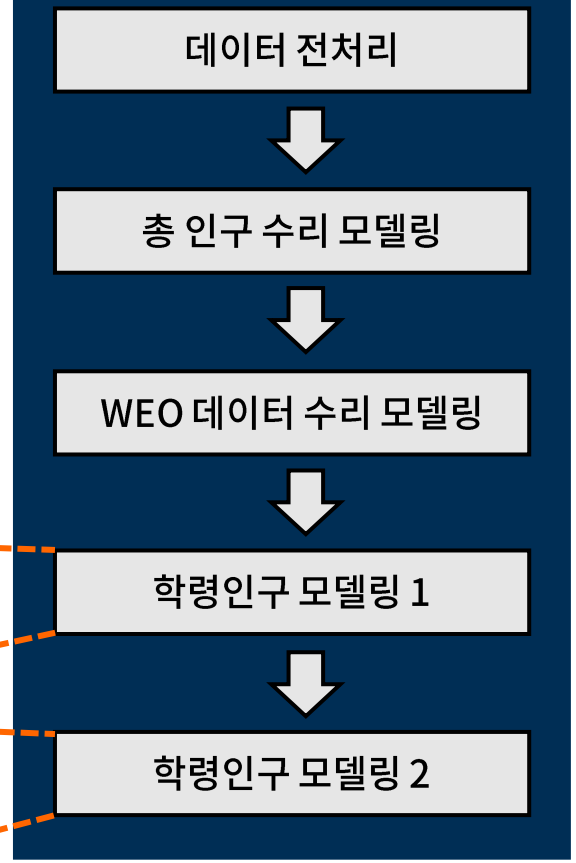
아이디어



활용 데이터

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

수행 과정



수리 모델링을 이용한 학령인구 예측
➔ 이차다변수다항함수

시계열 머신러닝 모델을 이용한 학령인구 예측
➔ SARIMA

02 프로젝트 수행과정

1 | 데이터 전처리 & 변수선택

공통) 데이터 슬라이싱

인구 데이터는 1960년부터의 데이터를, 경제 데이터는 1980년부터의 데이터를 가지고 있고 모두 2020년 기준이므로 그 이후의 값은 모두 예측 값

➡ 1980년부터 2020년까지의 실제 값인 데이터만 사용

인구데이터) 필요한 변수 추가

학령 인구에 해당하는 연령의 인구의 합인 '학령인구 수' 변수 추가, 학령인구 수를 총 인구 수로 나누어 '학령인구의 비율' 변수 추가

경제데이터) 변수 선택

WEO 데이터에 46개의 feature가 존재하여 상관관계가 높은 feature들이 다수 존재

➡ 아래 기준을 통해 2개의 feature를 선택

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

(* 인구 데이터: 연령별 인구 현황, 경제 데이터: WEO 데이터)

1-1 인구데이터 전처리

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

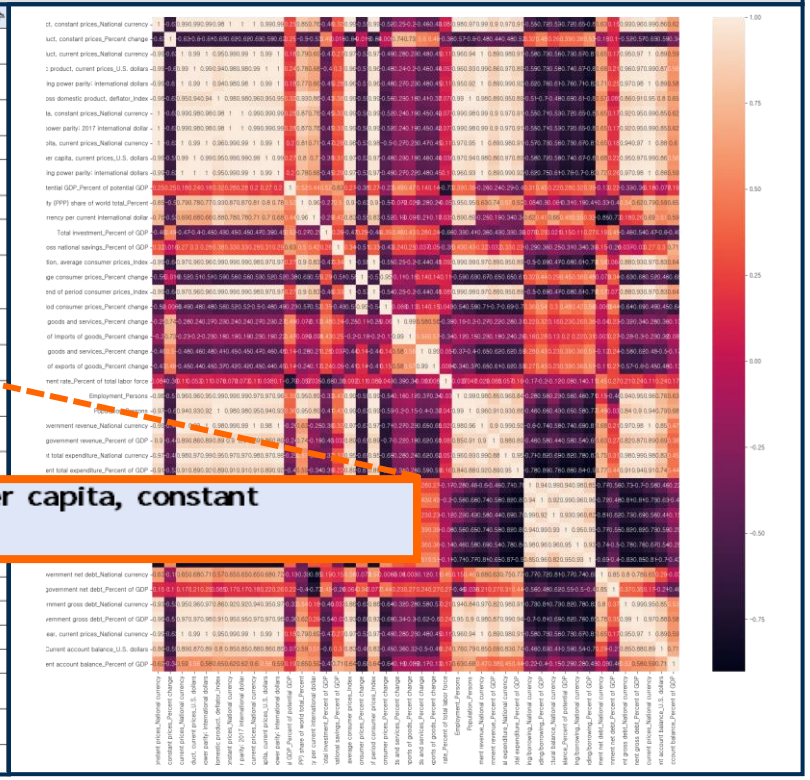
#	컬럼 명	데이터 개수	데이터 타입	결측치 개수
0	년도	111	int64	0
1	총인구수	111	int64	0
2	학령인구수	111	int64	0
3	학령인구비율	111	float64	0

02 프로젝트 수행과정

1-2 경제데이터 전처리

World Economic Outlook data (WEO) / 출처: 국제금융기구 (IMF)

#	컬럼 명	데이터 개수	#	컬럼 명	데이터 개수
0	Gross domestic product, constant prices_National currency	48	22	Volume of exports of goods and services_Percent change	48
1	Gross domestic product, constant prices_Percent change	48	23	Volume of exports of goods_Percent change	48
2	Gross domestic product, current prices_National currency	48	24	Unemployment rate_Percent of total labor force	48
3	Gross domestic product, current prices_U.S. dollars	48	25	Employment_Persons	44
4	Gross domestic product, current prices_Purchasing power parity; international dollars	48	26	Population_Persons	48
5	Gross domestic product, deflator_Index	48	27	General government revenue_National currency	33
6	Gross domestic product per capita, constant prices_National currency	48	28	General government revenue_Percent of GDP	33
7	Gross domestic product per capita, constant prices_Purchasing power parity; 2017 international dollar	48	29	General government total expenditure_National currency	33
8	Gross domestic product per capita, current prices_National currency	48	30	General government total expenditure_Percent of GDP	33
9	Gross domestic product per capita, current prices_U.S. dollars	48	31	General government net lending/borrowing_National currency	33
10	Gross domestic product per capita, current prices_Purchasing power parity; international dollars	48	32	General government net lending/borrowing_Percent of GDP	33
11	Output gap in percent of potential GDP_Percent of potential GDP	44	33	General government structural balance_National currency	33
	Gross domestic product based on		34	General government structural balance_Percent of pote	
			35	Gene lend	
			36	Gene lending/borrowing_Percent of GDP	
15	Gross national savings_Percent of GDP	48	37	General government net debt_National currency	25
16	Inflation, average consumer prices_Index	48	38	General government net debt_Percent of GDP	25
17	Inflation, average consumer prices_Percent change	48	39	General government gross debt_National currency	38
18	Inflation, end of period consumer prices_Index	48	40	General government gross debt_Percent of GDP	38
19	Inflation, end of period consumer prices_Percent change	48	41	Gross domestic product corresponding to fiscal year, current prices_National currency	48
20	Volume of imports of goods and services_Percent change	48	42	Current account balance_U.S. dollars	48
21	Volume of imports of goods_Percent change	48	43	Current account balance_Percent of GDP	48



Gross domestic product, current prices_Purchasing power parity; international dollars

Gross domestic product per capita, constant prices_National currency

Selected 경제 지표

#	컬럼 명	데이터 개수	데이터 타입	결측치 개수
0	GDP_global	47	float64	0
1	GDP_national	47	float64	0

02 프로젝트 수행과정

2 | 총 인구 예측 수리 모델링

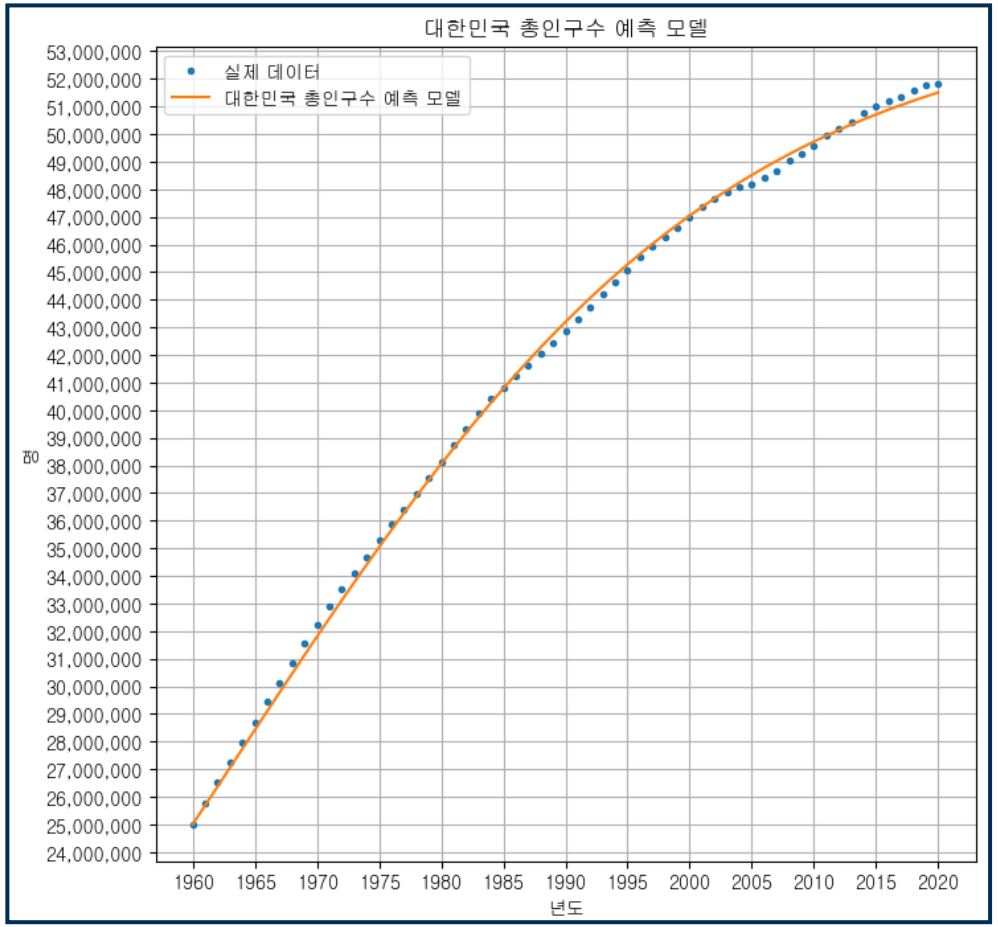
Population logistic equation (P.F. Verhulst)

$$\frac{dP}{dt} = P\left(r - \frac{r}{K}P\right)$$

1. 위의 인구 로지스틱 방정식을 이용하여 총 인구를 예측하는 수리 모델링을
2. 실제 데이터를 사용하여 최소제곱법 기법으로 파라미터를 추정하여 최적화

$$\frac{dP}{dt} = 2735514.20403647\left(\frac{e^{-0.0502031181087683t}}{(1+e^{-0.0502031181087683t})^2}\right)$$

우측의 그래프를 통해 실제 데이터와 비교하였을 때,
예측한 총 인구 수가 차이가 거의 나지 않는 것을 볼 수 있습니다.



⇒ 이 모델을 통해 2021년부터 2060년까지의 총 인구 수를 예측

02 프로젝트 수행과정

3 – 1 | WEO의 GDP_global 변수 예측

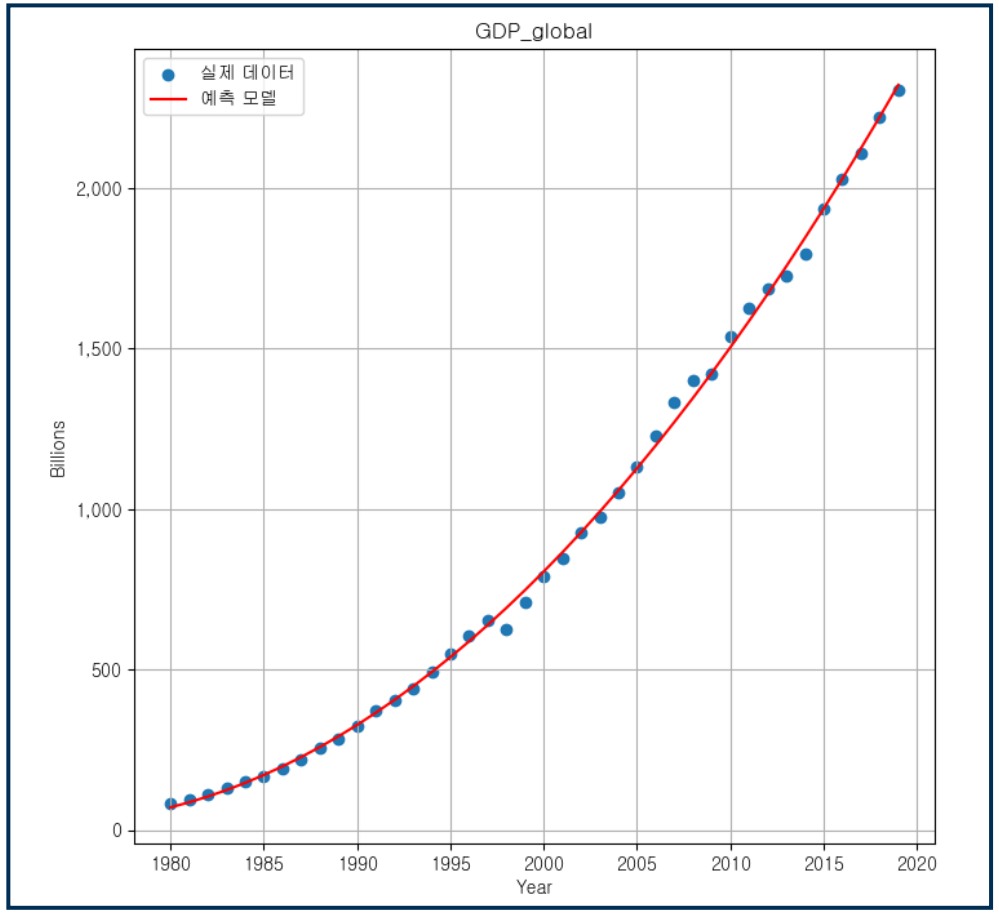
1. 기본적인 $Y = a\,t^2 + b\,t + c$ 이차함수를 이용하여 GDP_global을 예측하는 수식 모델링
2. 실제 데이터를 사용하여 최소제곱법 기법으로 파라미터를 추정하여 최적화

⇒ 최적화결과

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

우측의 그래프를 통해 수리 모델링을 통해 얻은 GDP_global 값이 실제 값과 유사한 분포를 가진 것을 볼 수 있습니다.

(* GDP_global: International dollars currency)



⇒ 이 모델을 통해 2021년부터 2060년까지의 GDP_global을 예측

02 프로젝트 수행과정

3 – 2 | WEO의 GDP_national 변수 예측

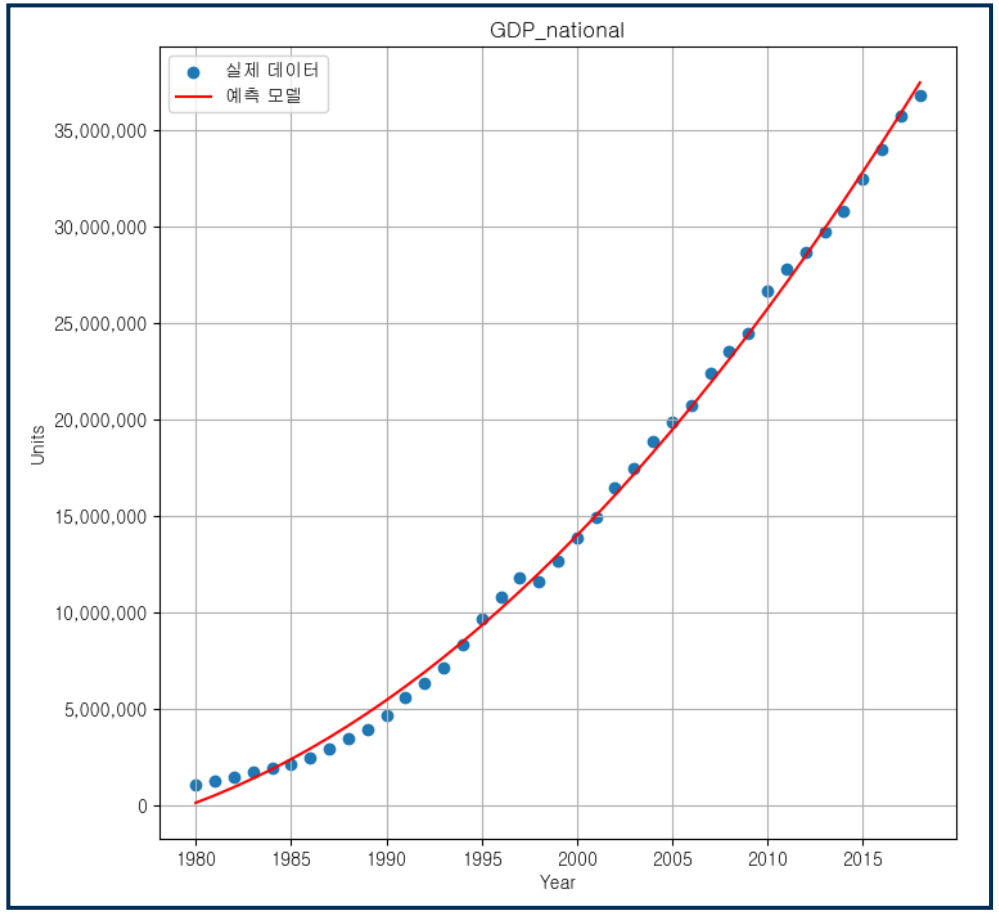
1. 기본적인 $Y = a\,t^2 + b\,t + c$ 이차함수를 이용하여 GDP_national을 예측하는 수식 모델링
2. 실제 데이터를 사용하여 최소제곱법 기법으로 파라미터를 추정하여 최적화

⇒ 최적화결과

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

우측의 그래프를 통해 수리 모델링을 통해 얻은 GDP_national 값이 실제 값과 유사한 분포를 가진 것을 볼 수 있습니다.

(* GDP_global: International dollars currency)



⇒ 이 모델을 통해 2021년부터 2060년까지의 GDP_national을 예측

02 프로젝트 수행과정

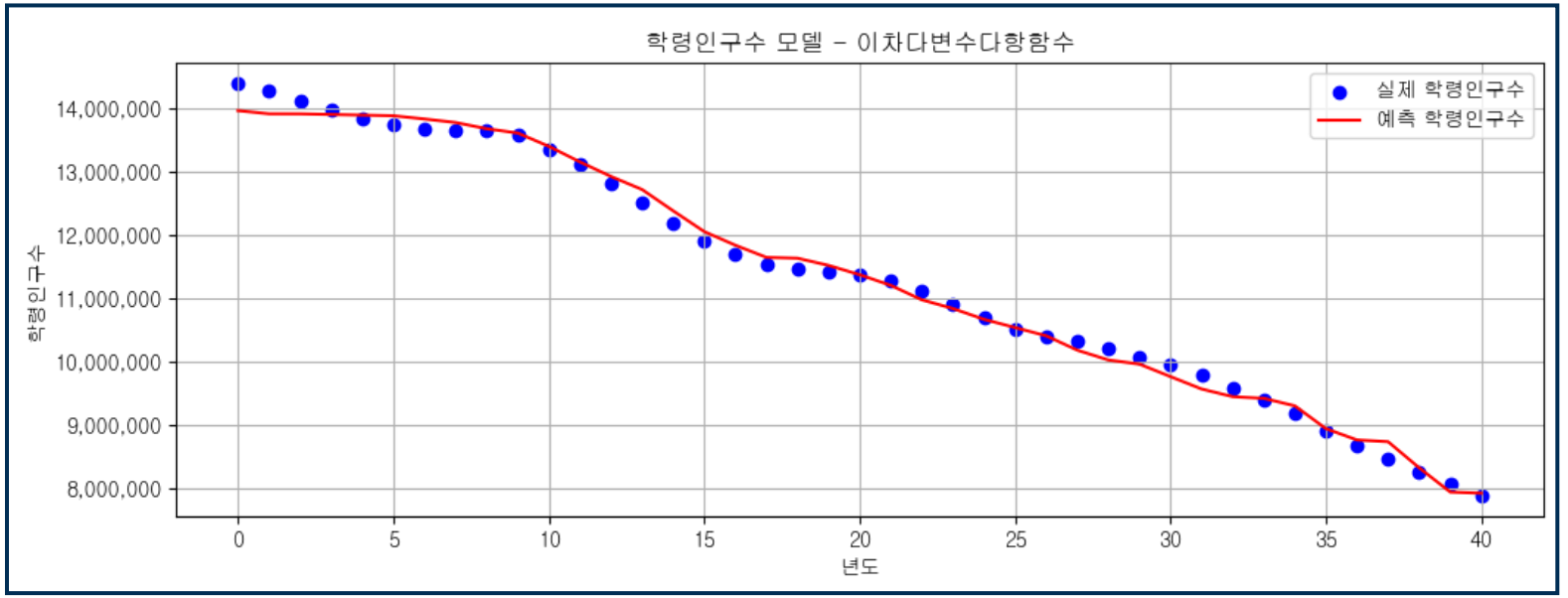
4 - 1 | 수리모델링을 이용한 학령인구 예측 with 경제데이터

1. 기본적인 $Y = a\,{x_1}^2 + b\,{x_2}^2 + cx_1x_2 + dx_1 + ex_2 + f$ 이차다변수다항함수를 이용하여 학령인구를 예측하는 수식 모델링
2. 실제 데이터를 사용하여 최소제곱법 기법으로 파라미터를 추정하여 최적화

⇒ 최적화결과

$$Y = 9.56592274\,{x_1}^2 + 5.92910228 \times 10^{-8}\,{x_2}^2 - 1.54072578 \times 10^{-3} x_1\,x_2 + 845.23441x_1 - 0.706889057x_2 + 14009144.2$$

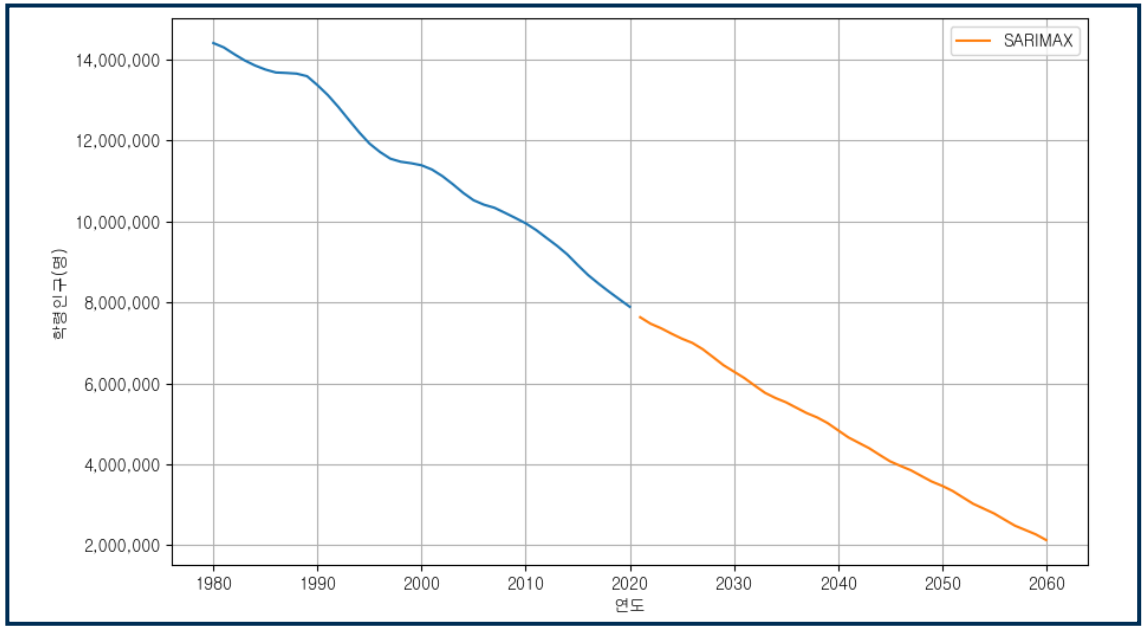
(Y: 학령인구, t: $year$, x_1: $GDP\ global$, x_2: $GDP\ national$)



02 프로젝트 수행과정

4 - 2 | 시계열모델을 이용한 학령인구 예측

사용모델	SARIMAX	
Non seasonal param	(p, d, q)	(2, 1, 0)
Seasonal param	(P, D, Q) [m]	(2, 1, 0) [4]
외생변수(exog)	총 인구 수	
AIC	-47.111	
BIC	-37.610	



1) 정규분포 검정

H0: 정규분포 O	H1: 정규분포 X

Jarque-Bera의 p-value: 1.00 >= 0.05
⇒ H0 채택, 즉 정규분포이다.

2) 이분산성 검정

H0: 등분산 O	H1: 등분산 X

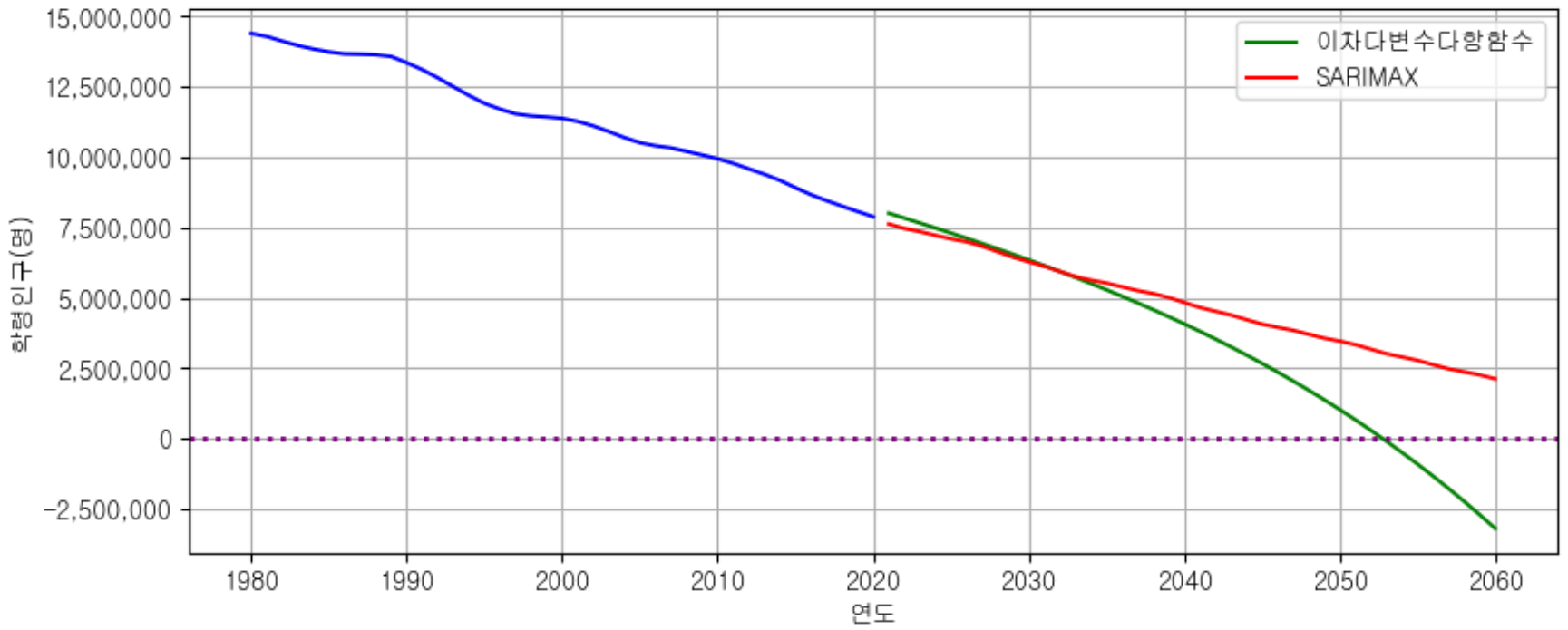
Heteroskedasticity의 p-value: 0.90 >= 0.05
⇒ H0 채택, 즉 등분산이다.

3) 자기상관성 검정

H0: 자기상관성 X	H1: 자기상관성 O

Ljung-Box의 p-value: 0.15 >= 0.05
⇒ H0 기각, but 이보다 좋은 값을 찾아내지 못함
따라서 85% 신뢰구간에서는 자기상관성이 있다

03 최종 모델 선정



최종 모델 선정 & 예측 값		
이차 다변수 다항함수	2030년 학령인구 예측	6,288,660
2050년대 초반, 학령인구가 0이하로 감소하며 추세 또한 너무 가파르게 감소 ➡ 기각	2040년 학령인구 예측	4,840,742
SARIMAX 모델	2050년 학령인구 예측	3,468,187
앞서 했던 모델 검증을 통하여 우수한 성능을 확인하였고, 감소하는 추세도 이전과 비슷하다고 판단 ➡ 최종 모델 선정	2060년 학령인구 예측	2,128,656

04 실현 방안

최종 모델의 장점

미래의 예측값인 독립변수에도 오차 범위가 존재하므로 독립변수가 많아지면 종속변수의 오차범위가 증가
But 최종 모델에는 비교적 오차범위가 적은 총인구수만을 독립변수로 채택하여 종속변수의 안정성을 확보

전체 인구 대비 특정 연령 구간의 비율을 구하는 방법 사용
➡ 학령인구 뿐만 아니라 영유아 인구, 노년 인구 등 특정 구간의 인구도 예측 가능
➡ 전체 학령인구 수 대비 특정 지역의 학령인구 비율예측에도 적용 가능

모델 활용 방안

특정 인구 세대를 타겟으로 하는 정책들 수립 시 활용 가능
ex) 국민연금 정책 수립을 위한 연금 수령 인구 및 경제활동 인구 예측

특정 지역의 학령인구를 타겟으로 하는 정책 수립시 활용 가능
ex) 부산지역의 학교 및 의료 시설 등 다양한 분야의 의사결정

참고문헌

데이터출처

* 행정안전부, “연령별 인구현황”, https://jumin.mois.go.kr/ageStatMonth.do, accessed 05 Nov 2023.
* IMF, “World Economic Outlook Database”, https://www.imf.org/en/Publications/WEO/weo-database/2021/April, accessed 05 Nov 2023.

참고문헌

* 송현정(2022), 2021년 출생 통계, 통계청
* 통계청(KOSIS), “주요 연령계층별 추계인구(생산연령인구, 고령인구 등) / 전국”, https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1BPA003&checkFlag=N, accessed 05 Nov 2023.
* 서울연구데이터서비스, “학령인구의 변화”, https://data.si.re.kr/data/%ED%86%B5%EA%B3%84%EB%A1%9C-%EB%B3%B8-%EC%84%9C%EC%9A%B8-%EC%9D%B8%EA%B5%AC%ED%8E%B8/272, accessed 05 Nov 2023.
* Dennis G.zill(2018), A First Course in DIFFERENTIAL EQUATIONS with Modeling Applications, 11th Edition, Cengage Learning

감사합니다