회귀분석을 통한
기대수명 예측 및 요인 분석
- 아시아 국가 중심으로 -

2020 년 6 월 26 일

동국대학교 이과대학 통계학과

201711**** 유 * *

2017111748 임정민

[목차]

Ⅰ. 서론

(1) 분석 배경 및 목적

(2) 분석 방법 및 절차

Ⅱ. 본론1

(1) 데이터 소개

(2) 데이터 전처리

Ⅲ. 본론2

(1) 전체 데이터 탐색

(2) 기대수명 예측 모델 구축 및 평가

Ⅳ. 본론3

(1) 아시아 국가 데이터 탐색

(2) 아시아 국가 기대수명에 영향을 주는 요인 분석

Ⅴ. 결론

(1) 분석 결과 해석

(2) 분석 한계 및 시사점

Ⅵ. 부록 (별도 파일 첨부)

Ⅰ 서론

(1) 분석 배경 및 목적

기대수명은 특정 연도의 0세 출생자가 생존할 것으로 기대되는 평균 생존 연수를 일컫는다.[1] 전 세계적으로 국가별 기대수명은 과거에도 꾸준히 연구되어온 주제이다. 질병이나 상해를 겪지 않는 건강수명 또한 관심사로 대두되고 있으며 국가별로 매년 기대수명의 증감에 영향을 끼친 요인은 무엇인지 다방면으로 조사되어왔다. 실제 대한민국 내 2018년 신생아의 기대수명은 82.7세였고, 소득에 따라 기대수명에 차이가 있다는 연구결과가 발표되었다.[2] 기대수명은 단순히 한사람이 몇 살까지 생존할 것인지 뿐만 아니라 한 국가의 사회적 문제가 무엇인지 나타내는 지표가 되기도 한다. 본 조는 회귀분석을 통해 전세계 국가들에 관한 정보를 바탕으로 각 국가의 기대수명을 예측하고 기대수명의 증감에 영향을 미치는 요인들에 대해 분석하고자 한다.

기대수명에 대한 관심이 커지며 기대수명 예측을 위한 양질의 데이터의 중요성이 강조되고 있다. 과거의 데이터는 1년동안 수집된 국가별 인구통계학적 변수와 소득과 사망률에 대한 정보만 포함하고 있었다. 본 조의 데이터는 한 국가의 기대수명에 영향을 줄 수 있는 B형간염, 소아마비, 홍역 등 다양한 질병변수 외에도 경제적변수, 사회적변수 등을 포함하고 있으며 2000 - 2015년 동안 수집된 데이터이므로 과거보다 더욱 세밀하게 기대수명을 예측할 수 있을 것으로 예상한다.

이러한 데이터를 바탕으로 진행하는 분석의 목적은 다음과 같다. 1) 선진국 여부에 따른 기대수명 예측 2) 아시아 국가의 기대수명에 영향을 주는 요인 분석

(2) 분석 방법 및 절차.

먼저 결측치, 이상치 등을 확인하여 데이터 전처리를 진행한다. 전처리를 마친 데이터 EDA를 통해 반응변수인 기대수명과 설명변수들 간의 관계를 파악한다. 이 과정에서 반응변수와 관계가 가장 높은 설명변수 1개를 선택해 단순선형회귀모형을 구축하고 평가한다. 이후, 변수선택법과 회귀모델 평가를 바탕으로 다중선형회귀모형을 구축하여 기대수명을 더욱 정확하게 예측한다.

또한, 우리나라가 속해 있는 아시아 대륙의 국가만을 대상으로 기대수명을 예측하는 두번째 회귀모델을 구축한다. 그 결과, 아시아 국가의 기대수명에 영향을 끼치는 요인이 무엇인지 확인하고 그 결과를 randomforest 모델의 변수 중요도 평가 결과와 비교한다.

Ⅱ 본론1

(1) 데이터 소개

WHO에 의해 수집된 193개국의 기대 수명, 건강 요인 데이터와 UN에 의해 수집된 각 나라의 경제 데이터로 구성되어 있다. 총 2938개의 행과 22개의 변수가 있으며, 사망률, 면역 관련 요인, 경제 및 사회적 요인으로 구분할 수 있다. 데이터 출처와 변수 설명은 [부록1]에 제시되어 있다.

---

[1] [NAVER 지식백과] "기대수명" (통계표준용어, 통계청)

[2] 김명희(2019), "포용복지와 건강정책의 방향", 보건복지포럼, 278 (0), 한국보건사회연구원, 30-43

(2) 데이터 전처리

데이터에서 각 변수에 존재하는 행의 개수가 다른 것을 보면 결측치가 존재함을 알 수 있다. 전처리 단계에서 결측치를 대체하기 전에 각 변수의 특성을 고려했을 때 잘못된 값이 존재하지 않는지 확인할 필요가 있다.

① 변수 탐색

Country와 Year, Status 변수를 제외하고 나머지 19개의 변수들의 기술통계량을 살펴보자. 아래의 표는 문제가 있어 보이는 변수들만 보여준다.

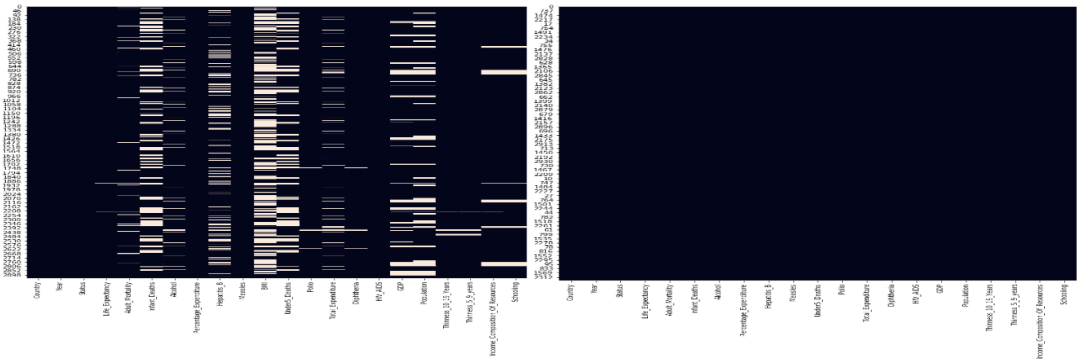
	Adult_Mortality	Infant_Deaths	Under5_Deaths	BMI
min	1	0	0	1
25%	74	0	0	19.3
50%	144	3	4	43.5
75%	228	22	28	56.2
max	723	1800	2500	87.3

Adult_Mortality는 성인의 사망률을 의미하는 변수인데, 변수의 최솟값이 1인 것은 15세에서 60세 사이에 사망할 가능성이 0.1%인 것을 의미하므로 문제가 있어 보인다. 마찬가지로 Infant_Deaths와 Under5_Death는 유아 사망률을 의미하는 변수인데, 최솟값과 25% 값이 0인 것은 오류일 것으로 생각된다. 마지막으로 BMI 변수의 특성을 살펴보면, 15보다 작으면 저체중, 40보다 크면 고도비만이다. 이와 같은 기준에 따르면 BMI 변수의 최솟값과 최댓값은 잘못 입력된 값으로 생각할 수 있다. 이들 변수 각각 기준을 세워서 NULL값으로 대체한다.

	기준	대체값
Adult_Mortality	x < 5 percentile	NULL
Infant_Deaths	x = 0	NULL
Under5_Deaths	x = 0	NULL
BMI	x < 10 or x > 50	NULL

② 결측치 처리

각 변수에 존재하는 결측치의 개수와 비율을 보여주는 표는 [부록2]에 제시되어 있다. 먼저, 결측치의 비율이 거의 50%인 BMI 변수는 제거한다. 나머지 변수들은 결측치를 다른 값으로 대체한다. 대체 방법 중 하나는 국가 별로 보간(interpolation)하는 것이다. 그러나 어떤 특정 국가에만 결측이 존재하는 경우가 많이 있기 때문에 결측치를 대체할 수 없는 상황이 발생한다. 따라서 연도 별로 보간하는 방법을 이용한다. 즉, 연도별로 각 변수의 중앙값을 결측치 대체값으로 사용한다. 그 결과 아래의 그림처럼 존재하던 결측치들이 모두 사라진 것을 볼 수 있다.



③ 이상치 처리

각 변수의 boxplot과 histogram을 보고 이상치 존재를 파악한다. [부록3]에 있는 그래프를 보면, 반응변수 Life_Expectancy 변수 외에도 여러 설명변수에 이상치가 존재하는 것을 볼 수 있다. 또한, [부록4]에 있는 표는 2xIQR을 기준으로 각 변수의 이상치 개수와 비율을 보여준다

각 변수에 존재하는 이상치를 윈저화 방법(winsorization)을 이용하여 대체한다. 이상치를 제외한 나머지 값 중에서 최댓값 또는 최솟값에 가까운 값으로 이상치를 대체하는 방법이다. 이상치가 처리된 변수들의 boxplot 그래프는 [부록5]에서 보여준다.

III 본론2

(1) 전체 데이터 탐색

전체 데이터를 7:3 비율로 분할하여 학습용 데이터를 탐색한다. 데이터 탐색을 통해 기대수명에 영향을 끼치는 변수를 알아보고 설명변수 사이의 상관성을 확인한다.

① 변수 분포

Country와 Year, Status 변수를 제외하고 나머지 18개 변수들의 분포를 살펴보기 위해 각 변수들의 히스토그램을 [부록6]에서 제시한다. 또한, 선진국과 개발도상국으로 나누어지는 Status 변수의 분포 비율을 보여주는 그래프는 [부록7]에서 제시한다.

② Status에 따른 기대수명 차이

[부록8]에 제시된 그래프와 오른쪽 표를 보면, Status 변수에 따라 기대수명에 유의미한 차이가 있다고 볼 수 있다. 선진국과 개발도상국 집단 간에 기대 수명 평균의 차이가 있는지 알아보기 위해 T-test를 진행한다.

Status	mean
Developed	79.197852
Developing	67.157172

T-test 결과, t 통계량은 47.9 이고 P-value는 6e-323으로 유의수준 0.05보다 작다. 즉, 두 집단 간에 기대 수명 평균의 차이가 있다고 할 수 있다. 따라서 모델 구축 단계에서 선진국과 개발도상국을 의미하는 Status 범주형 변수를 회귀모델 설명변수로 사용한다.

③ 상관관계

전체 데이터의 상관관계를 보여주는 heatmap 그래프가 [부록9]에 제시되어 있다. 반응변수인 기대수명과 가장 높은 상관관계를 가지는 변수는 HIV_AIDS 와 Income_Composition_Resources 이다. HIV_AIDS 변수의 상관계수는 -0.8이고 Income_Composition_Resources 변수의 상관계수는 +0.8이다. 따라서 단순선형회귀를 구축할 때 HIV_AIDS 변수 또는 Income_Composition_Resources 변수를 설명변수로 사용한다. 또한, 기대수명 변수를 제외한 나머지 변수들 사이에 공선성이 존재하는 것을 heatmap 그래프로 예상할 수 있다. 따라서 모델 구축 단계마다 VIF 값으로 다중 공선성을 확인한다.

(2) 기대수명 예측 모델 구축 및 평가

① 상수항만 있는 회귀 모델

반응변수를 기대수명으로 하는 회귀모델을 구축한다. 변수선택법을 이용하여 다중선형회귀모델로 확장하기 위해서 우선 상수항만 있는 모델을 구축한다. 결과는 아래의 표와 같다.

	Coef	Std err	t	P > |t|
Intercept	69.1645	0.211	327.929	0.000

Adj.R²	Prob (F-statistic)
0.000	nan

추정된 회귀식은 $\hat{y} = 69.2$ 이고 조정된 R-squared값은 0이다. 즉, 전체국가의 기대수명 평균값은 69.2세이다. 이제 변수를 추가하여 단순선형회귀모델을 만들고 다중선형회귀로 확장한다.

② 단순선형회귀모델

반응변수인 기대수명과 가장 큰 상관관계에 있는 HIV_AIDS 변수를 사용하여 단순선형회귀모델을 구축한다. 결과는 아래와 같다.

	Coef	Std err	t	P > |t|
Intercept	75.1235	0.164	459.036	0.000
HIV_AIDS	-11.2000	0.190	-58.905	0.000

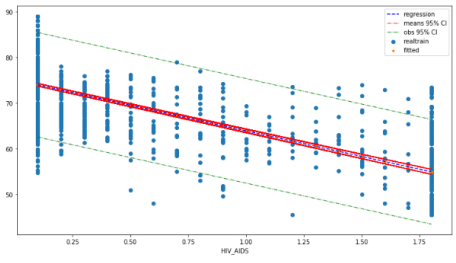
Adj.R²	Prob (F-statistic)
0.628	0.00

추정된 회귀식은 $\hat{y} = 75.1 - 11.2 * HIV\_AIDS$ 이고 조정된 R-squared 값은 0.628이다. 즉, 각 나라의 에이즈 전염병으로 이한 사망률에 따라 기대수명이 변화한다. 단순 회귀로 해석한 결과, 에이즈 전염병으로 인한 사망률이 0인 국가의 기대수명은 75.1세이고 사망률이 1만큼 증가할수록 기대수명이 평균적으로 11.2세만큼 감소한다. 회귀 직선 구축 결과 summary 표는 [부록10]에서 보여준다.

다중선형회귀모델로 확장하기 전에 먼저 단순회귀모델이 가정을 만족하는지 확인한다. 설명변수가 1개인 모델이기 때문에 공선성 가정을 제외하고 선형성, 잔차의 정규성, 잔차의 등분산성, 잔차의 독립성 가정을 만족하는지 확인한다.

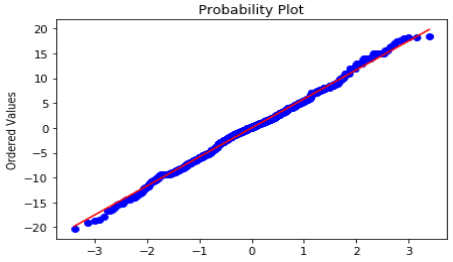
▷ 선형성

오른쪽 그래프는 반응변수와 설명변수의 선형성을 확인하기 위해 산점도 그래프에 구축된 회귀 직선을 시각화 하였다. 두 변수 사이의 상관계수는 -0.8이다. 따라서 회귀모델이 선형성 가정을 만족한다고 볼 수 있다.



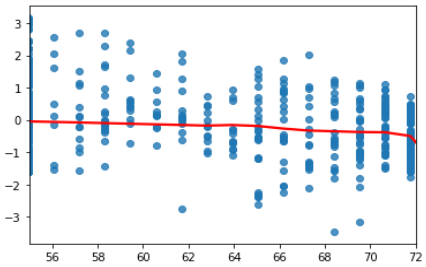
▷ 잔차의 정규성

오른쪽 그래프는 회귀모델의 잔차가 정규성을 만족하는지 확인하기 위해 그린 Q-Q Plot이다. 점이 찍힌 형태가 빨간색 직선에 근접함을 볼 수 있다. 따라서 회귀 모델의 잔차가 정규성을 따른다고 볼 수 있다.



▷잔차의 등분산성

오른쪽 그래프는 회귀모델의 잔차가 등분산성을 만족하는지 확인하기 위해 그린 그래프이다. 빨간 선은 잔차의 트렌드를 나타내는 선으로, 이 선이 수평에 가까운 것을 보아 잔차가 등분산성을 따른다고 볼 수 있다.



▷잔차의 독립성

모델의 잔차의 독립성 가정을 따르는지 확인하기 위해 시행한 더빈왓슨 검정 결과는 오른쪽 표와 같다. 그 결과, 유의확률이 0.05보다 크기 때문에 단순선형회귀모델이 잔차의 독립성을 만족한다는 것을 확인할 수 있다.

Durbin Watson Test	
DW	1.9678
P-Value	0.2322

③ 다중선형회귀모델

가정을 만족하는 단순선형회귀모델에 설명 변수를 추가하여 다중회귀모델로 확장한다. 이때, 변수를 추가하는 과정에서 단계적 변수 선택법을 사용한다. 즉, 모델에 사용되지 않은 변수 중에서 기준 통계치를 개선시키는 변수를 추가하거나 모델에 사용된 변수 중에서 유의하지 않은 변수를 제거한다. 또한 변수를 추가할 때마다 VIF값으로 다중 공선성을 진단한다. 아래의 표는 최종 모델에 사용된 변수 및 추정된 회귀식의 기준 통계치 변화를 각 단계마다 보여준다. 자세한 변수 추가 및 제거 과정은 [부록11]에서 제시한다.

단계	변수	Adj.R-squared	AIC	BIC
1단계	Status	0.701	1.264e+04	1.266e+04
2단계	Income	0.799	1.182 e+04	1.185 e+04
3단계	Adult_Mortality	0.824	1.155 e+04	1.158 e+04
4단계	Diphtheria	0.835	1.143 e+04	1.146 e+04
5단계	Infant_Deaths	0.844	1.130 e+04	1.134 e+04
6단계	GDP	0.846	1.128 e+04	1.133 e+04
7단계	Thinness_5_9_years	0.849	1.124 e+04	1.130 e+04
8단계	Measles	0.853	1.119 e+04	1.125 e+04

각 단계마다 결정계수 값이 증가하고 AIC, BIC값이 감소하는 것을 볼 수 있다. 한편, 최종 모델에는 2개의 교호작용항이 포함된다. 변수 추가 단계마다 Status와의 교호작용이 유의한지 그리고 full model이 더욱 효과적인지 F-test를 시행한 결과, 최종 모델에는 Thinness_5_9_years 변수와 Status 변수 그리고 Measles 변수와 Status변수의 교호작용이 추가되었다. 따라서 최종 다중회귀모형 구축 결과는 아래의 표와 같다. 다중회귀직선 구축 결과 summary 표는 [부록12]에서 보여준다.

	비표준화계수	표준화계수	Std err	P > |t|
Intercept	63.6379		0.879	0.000
HIV_AIDS	−5.4020	−0.3824	0.167	0.000
Status	−3.9478	−0.1568	0.455	0.000
Income_Composition_Of_Resources	16.3900	0.2932	0.774	0.000

Adult_Mortality	−0.0154	−0.1739	0.001	0.000
Diphtheria	0.0602	0.0963	0.007	0.000
Infant_Deaths	−0.0275	−0.0739	0.004	0.000
GDP	7.073e−05	0.0392	1.97e−05	0.000
Thinnes_5_9_years	−1.8869	−0.7821	0.254	0.000
Measles	0.0021	0.0738	0.001	0.014
Status : Thinness_5_9_years	1.7738	0.1256	0.255	0.000
Status : Measles	−0.0035	−0.0001	0.001	0.000

Adj.R^2	Prob (F-statistic)	AIC	BIC
0.854	0.00	1.118e+04	1.125e+04

위 결과를 보면, 기대수명에 대한 최종 모델의 분산 설명력은 85.4%이고 회귀모형은 유의미한 것으로 나타난다. 기대수명에 큰 영향을 끼치는 요인을 확인하기 위해 표준화 변수를 확인하자. 9개의 설명변수 중에서 Thinness_5_9_years, HIV_AIDS, Income_Composition_Of_Resources, Adult_Mortality 그리고 Status 순으로 기대수명에 영향을 끼친다.

먼저, 기대수명에 가장 큰 영향을 끼치는 요인은 Thinness_5_9_years 변수이다. 이 변수는 저체중 어린이 비율을 뜻한다. 나머지 설명변수들이 상수로 고정되었을 때, 저체중인 어린이 비율(%)이 증가할수록 기대수명은 감소한다. 즉, Thinness_5_9_years 변수가 1% 증가할 때 해당 국가의 기대수명은 약 2년 감소한다.

단순선형회귀모델의 설명변수인 HIV_AIDS 변수도 기대수명에 큰 영향을 끼친다. 이 변수는 AIDS로 사망한 유아의 비율을 뜻하는데, permil 단위가 사용되고 있다. 따라서 나머지 설명변수들이 상수로 고정되었을 때, AIDS로 사망한 유아의 비율이 0.1%(=1 permil) 증가하면 기대수명이 약 5년 감소한다고 볼 수 있다.

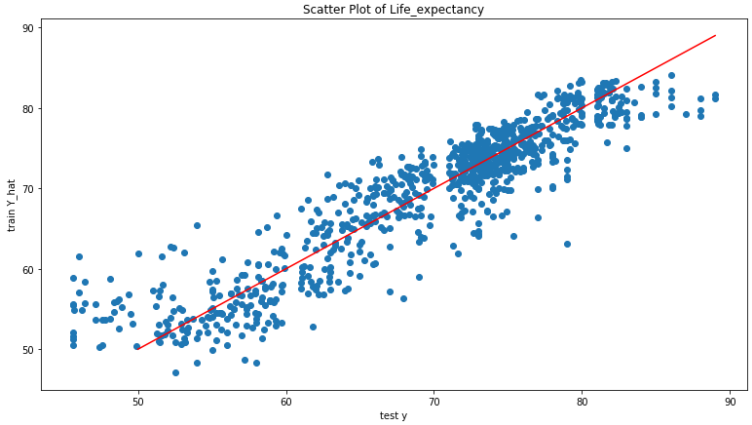
또한, 기대수명과 높은 상관관계에 있는 Income_Composition_Of_Resources 변수는 기대수명 증가에 영향을 끼친다. 즉, HRDI 값이 높을수록 기대수명이 크게 증가한다고 볼 수 있다. 따라서 나머지 설명변수들이 상수로 고정되었을 때, HRDI 값이 10%(=0.1) 증가하면 해당 국가의 기대수명이 약 2년 증가한다.

마지막으로 기대수명과 -0.7 상관계수를 가지는 Adult_Mortality 변수는 기대수명 감소에 영향을 끼친다. 이 변수는 성인의 사망률을 뜻하는데, permil 단위가 사용되고 있다. 따라서 나머지 설명변수들이 상수로 고정되었을 때, 성인의 사망률이 10%(=100permil) 증가하면 기대수명이 약 1.5년 감소한다고 볼 수 있다.

한편, 데이터 탐색 단계에서 확인한 바와 같이 선진국과 개발도상국 간의 기대수명 차이가 있다. 나머지 설명변수들이 고정되어 있을 때, 선진국의 기대수명은 개발도상국보다 약 4년이 더 길다.

선진국 여부 변수와의 교호작용항을 살펴보자. 선진국 여부에 따라 5세에서 9세 사이의 저체중 비율이 기대수명에 끼치는 영향이 변화한다. 즉, 개발도상국가일 때는 아이들의 저체중 비율이 1% 증가하면 해당 국가의 기대수명은 약 4년 감소하고 선진국에서는 약 2년 감소한다. 그리고 선진국 여부에 따라 홍역 발병 비율이 기대수명에 끼치는 영향도 변화하는데, 개발도상국가일 때는 홍역 발병 비율이 0.1%(=1permil) 증가하면 기대수명이 약 4년 감소하고 선진국에서는 기대수명에 큰 영향을 끼치지 않는다.

아래 그래프는 최종 회귀 모델의 예측력을 확인하기 위한 자료이다. 산점도 그래프가 y=x 빨간 직선에 가까울수록 모델이 예측한 y값과 실제 test 데이터의 y값이 일치함을 의미한다.



이제 최종 모델이 가정을 만족하는지 확인하자. 최종 모델이 다중회귀모델이기 때문에 독립성 가정을 포함하여 선형성, 잔차의 정규성, 잔차의 등분산성 잔차의 독립성 가정을 모두 만족하는지 확인한다.

▷ 선형성

아래의 표는 각 설명변수가 기대수명과 선형관계에 있는지 확인하기 위해서 반응변수와 설명변수 사이의 상관계수를 보여준다.

features	Correlation	features	Correlation
HIV_AIDS	-0.8	Infant_Deaths	-0.5
Status	-0.5	GDP	0.5
Income_Composition_Of_Resources	0.8	Thinnes_5_9_years	-0.5
Adult_Mortality	-0.7	Measles	0.4
Diphtheria	0.6		

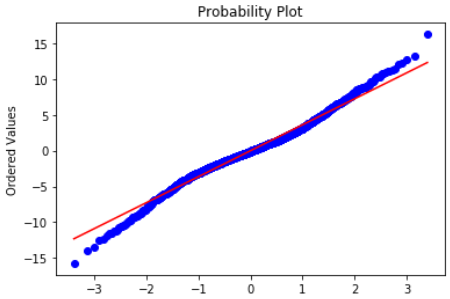
▷ 독립성

설명변수들 간의 독립성을 확인하기 위해 모델의 다중 공선성을 진단하는 VIF지수를 이용한다. 아래의 표는 각 설명변수의 VIF값을 보여준다. 모든 설명변수의 VIF값이 5를 넘지 않는 작은 값이기 때문에 다중 공선성이 발생할 위험이 없다. 즉, 설명변수 간의 상관관계가 약하기 때문에 독립성 가정을 만족한다고 볼 수 있다.

features	VIF Factor	features	VIF Factor
HIV_AIDS	1.966	Infant_Deaths	1.731
Status	1.499	GDP	1.638
Income_Composition_Of_Resources	2.694	Thinnes_5_9_years	1.529
Adult_Mortality	1.663	Measles	1.625
Diphtheria	1.514		

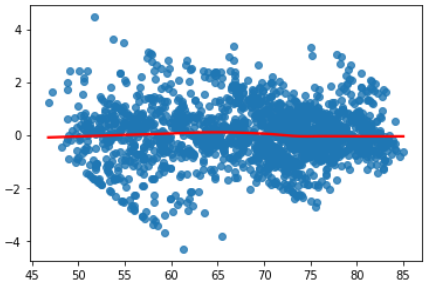
▷ 잔차의 정규성

오른쪽 그래프는 회귀모델의 잔차가 정규성을 만족하는지 확인하기 위해 그린 Q-Q Plot이다. 점이 찍힌 형태가 빨간색 직선에 근접함을 볼 수 있다. 따라서 회귀 모델의 잔차가 정규성을 따른다고 볼 수 있다.



▷잔차의 등분산성

오른쪽 그래프는 회귀모델의 잔차가 등분산성을 만족하는지 확인하기 위해 그린 그래프이다. 빨간 선은 잔차의 트렌드를 나타내는 선으로, 이 선이 수평에 가까운 것을 보아 잔차가 등분산성을 따른다고 볼 수 있다.



▷잔차의 독립성

모델의 잔차의 독립성 가정을 따르는지 확인하기 위해 시행한 더빈왓슨 검정 결과는 오른쪽 표와 같다. 그 결과, 유의확률이 0.05보다 크기 때문에 단순선형회귀모델이 잔차의 독립성을 만족한다는 것을 확인할 수 있다.

Durbin Watson Test	
DW	2.0136
P-Value	0.6209

Ⅳ. 본론3

(1) 아시아 국가 데이터 탐색

전체 데이터에서 아시아 국가에 해당하는 데이터를 추출하여 아시아 대륙의 기대수명에 영향을 끼치는 요인을 분석하고자 한다. 이때, 기대수명 예측이 아닌 요인분석이 목적이므로 학습용 데이터와 테스트 데이터를 분리하지 않고 전체 데이터를 사용하여 분석을 진행한다.

① Status에 따른 기대수명 차이

[부록13]에 제시된 그래프와 오른쪽 표를 보면, 아시아 대륙에서도 Status 변수에 따라 기대수명에 유의미한 차이가 있다고 볼 수 있다. 선진국과 개발도상국 집단 간에 기대 수명 평균의 차이가 있는지 알아보기 위해 T-test를 진행한다.

Status	mean
Developed	82.006250
Developing	70.158333

T-test결과, t 통계량은 32.5 이고, P-value는 6e-43으로 유의수준 0.05보다 작다. 즉, 두 집단 간에 기대 수명 평균의 차이가 있다고 할 수 있다. 따라서 모델 구축 단계에서 선진국과 개발도상국을 의미하는 Status 범주형 변수를 회귀모델 설명변수로 사용한다.

② 상관관계

아시아 데이터의 상관관계를 보여주는 heatmap 그래프가 [부록14]에 제시되어 있다. 반응변수인 기대수명과 높은 상관관계를 가지는 변수는 Income_Composition_Of_Resources 변수와 Schooling 변수 그리고 Adult_Mortality 변수이다. 이 중에서 상관계수가 0.7인 Schooling 변수를 단순선형회귀의 설명변수로 사용한다. 뿐만 아니라, 기대수명 변수를 제외한 나머지 변수들 사이에 공선성이 존재하는 것을 heatmap그래프로 예상할 수 있다. 따라서 모델 구축단계마다 VIF 값으로 다중 공선성을 확인한다.

(2) 기대수명 예측 모델 구축 및 평가

① 상수항만 있는 회귀 모델

반응변수를 기대수명으로 하는 회귀모델을 구축한다. 변수선택법을 이용하여 다중선형회귀모델로 확장하기 위해서 우선 상수항만 있는 모델을 구축한다. 결과는 아래의 표와 같다.

	Coef	Std err	t	P > |t|
Intercept	70.7363	0.222	318.882	0.000

Adj.R^2	Prob (F-statistic)
0.000	nan

추정된 회귀식은 $\hat{y} = 70.7$ 이고 조정된 R-squared 값은 0이다. 즉, 전체 국가의 기대수명 평균값은 70.7세이다. 이제 변수를 추가하여 단순선형회귀모델을 만들고 다중선형회귀로 확장한다.

② 단순선형회귀모델

반응변수인 기대수명과 큰 상관관계에 있는 Schooling 변수를 사용하여 단순선형회귀모델을 구축한다. 결과는 다음과 같다.

	Coef	Std err	t	P > |t|
Intercept	48.2450	0.877	54.995	0.000
Schooling	1.9274	0.074	26.050	0.000

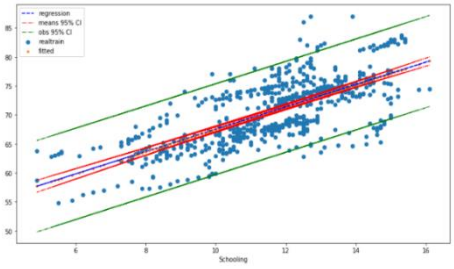
Adj.R^2	Prob (F-statistic)
0.508	0.00

추정된 회귀식은 $\hat{y} = 48.2 + 1.9 * Schooling$ 이고 조정된 R-squared 값은 0.508이다. 즉, 각 나라의 교육 기간에 따라 기대수명이 변화한다. 단순 회귀로 해석한 결과, 학업을 이수한 기간이 길수록 기대수명이 평균적으로 1.9세만큼 증가한다. 자세한 회귀 직선 구축 결과 summary는 [부록15]에서 보여준다.

다중선형회귀모델로 확장하기 전에 먼저 단순회귀모델이 가정을 만족하는지 확인한다. 설명변수가 1개인 모델이기 때문에 공선성 가정을 제외하고 선형성, 잔차의 정규성, 잔차의 등분산성, 잔차의 독립성 가정을 만족하는지 확인한다.

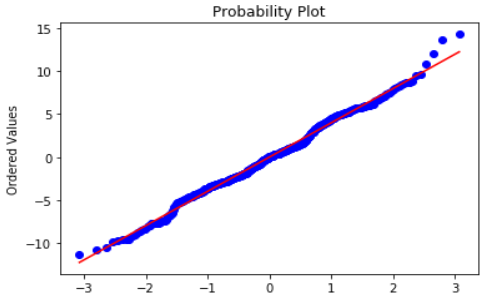
▷ 선형성

오른쪽 그래프는 산점도 그래프에 구축된 회귀 직선을 시각화한 자료이다. 두 변수 사이의 상관계수는 0.7이다. 따라서 회귀모델이 선형성 가정을 만족한다고 볼 수 있다.



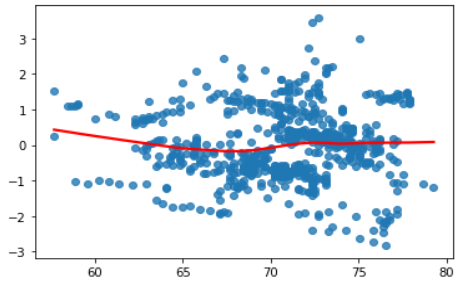
▷ 잔차의 정규성

오른쪽 그래프는 회귀모델의 잔차가 정규성을 만족하는지 확인하기 위해 그린 Q-Q Plot이다. 점이 찍힌 형태가 빨간색 직선이 근접함을 볼 수 있다. 따라서 회귀모델의 잔차가 정규성을 따른다고 볼 수 있다.



▷잔차의 등분산성

오른쪽 그래프는 회귀모델의 잔차가 등분산성을 만족하는지 확인하기 위해 그린 그래프이다. 빨간 선은 잔차의 경향성을 나타내는 선으로, 이 선이 수평에 가까운 것을 보아 잔차가 등분산성을 따른다고 볼 수 있다.



▷잔차의 독립성

모델의 잔차의 독립성 가정을 따르는지 확인하기 위해 시행한 더빈왓슨 검정 결과는 오른쪽 표와 같다. 그 결과, 유의확률이 0.05보다 크기 때문에 단순선형회귀모델이 잔차의 독립성을 만족한다는 것을 확인할 수 있다.

Durbin Watson Test	
DW	2.0723
P-Value	0.8209

③ 다중선형회귀모델

가정을 만족하는 단순선형회귀모델에 설명 변수를 추가하여 다중회귀로 확장한다. 이때, 변수를 추가하는 과정에서 앞선 본론2의 회귀분석 절차와 마찬가지로 단계적 변수 선택법을 사용한다. 아래의 표는 최종 모델에 사용된 변수 및 추정된 회귀식의 기준 통계치 변화를 각 단계마다 보여준다.

단계	변수	Adj.R-squared	AIC	BIC
1단계	Adult_Mortality	0.665	3426	3440
2단계	GDP	0.679	3398	3416
3단계	Polio	0.699	3359	3382
4단계	Income_Composition_Of_Resources	0.726	3299	3326
5단계	Diphtheria	0.730	3290	3321
6단계	Polio (제거)	0.730	3289	3316
7단계	Status	0.756	3224	3255
8단계	GDP (제거)	0.756	3223	3249
9단계	Alcohol	0.762	3207	3238
10단계	HIV_AIDS	0.764	3203	3239

위의 표를 통해 각 단계마다 변수를 추가, 제거함에 따라 결정계수가 증가하고 AIC, BIC 값이 감소하는 것을 볼 수 있다. 즉, 새로운 변수가 추가하고 유의하지 않은 변수를 제거하는 과정을 거치며 결정계수와 AIC, BIC값이 개선됨을 확인할 수 있다. 자세한 변수 추가 또는 제거 과정은

[부록16]에서 보여준다.

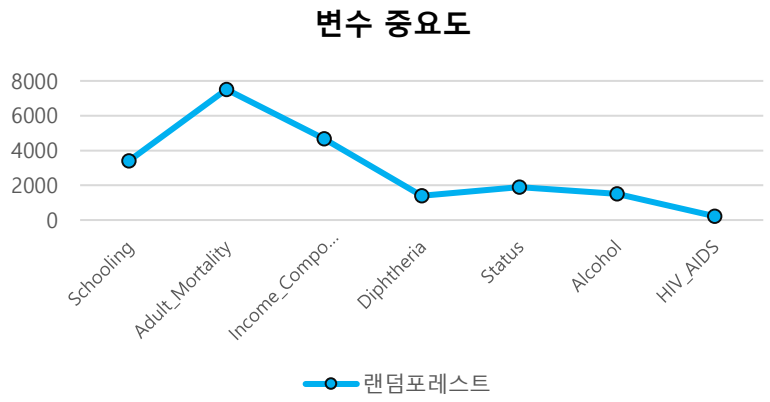
한편, Status에 따른 기대수명의 차이가 큰 것을 앞서 T test를 통해 확인한 바 있다. 그 점에 주목해 각 설명변수와 Status변수의 교호작용항을 적용해 보았다. 그 결과, 모델에서 교호작용항은 모두 유의하지 않았다. 따라서 최종 다중회귀모형 구축 결과는 아래의 표와 같다. 자세한 회귀 직선 구축 결과 summary는 [부록17]에서 보여준다.

	비표준화계수	표준화계수	Std err	P > |t|
Intercept	60.5934		1.175	0.000
Schooling	0.4796	0.1767	0.088	0.000
Adult_Mortality	−0.0281	−0.3604	0.002	0.000
Income_Composition_Of_Resources	10.6569	0.2646	1.310	0.000
Diphtheria	0.0667	01608	0.010	0.000
Status	−4.5740	−0.1735	0.547	0.000
Alcohol	0.2193	0.0906	0.051	0.000
HIV_AIDS	−1.4520	−0.0483	0.593	0.015

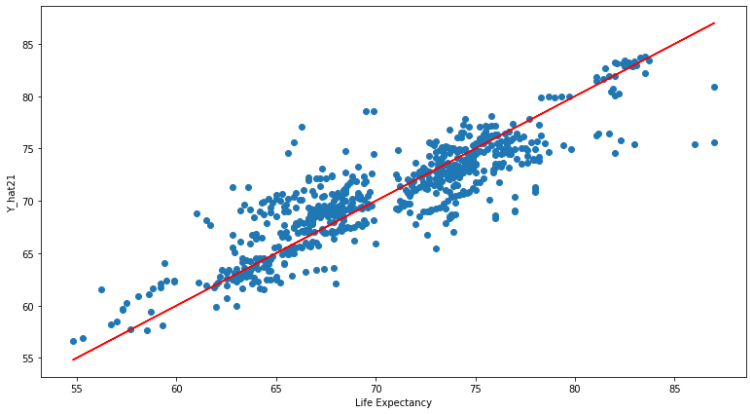
Adj.R^2	Prob (F-statistic)	AIC	BIC
0.764	0.00	3203	3239

위의 결과를 보면, 기대수명에 대한 최종 모델의 분산 설명력은 76.4% 이고 회귀모형은 유의미한 것으로나타난다. 하지만 각 변수의 단위가 모두 달라서 비표준화 계수를 통해 기대수명에 영향을 미치는 요인을 논하는 것은 무리가 있다. 따라서 기대수명의 요인분석을 위해 표준화 계수로 반응변수에 큰 영향을 미치는 변수를 파악한다.

표준화 계수의 절댓값을 확인한 결과, Adult_Mortality, Income_ Composition_Of_Resources, Schooling, Status 순으로 기대수명에 중요한 영향을 끼치고 있음을 확인할 수 있다. 이러한 결과가 랜덤포레스트 변수 중요도 평가와 일치하는지 확인한다. 오른쪽 그래프는 랜덤포레스트 모형에서의 변수 중요도를 보여준다. 그 결과, 랜덤포레스트 모형에서의 변수 중요도와 다중선형회귀모형에서의 변수 중요도가 대략적으로 일치하는 것을 알 수 있다.



아래는 최종 회귀 모델의 예측력을 확인하기 위한 그래프이다. 산점도 그래프가 y=x 직선에 가까울수록 모델이 예측한 y값과 실제 데이터의 y값이 일치함을 의미한다.



이제 최종 모델이 가정을 만족하는지 확인하자. 최종 모델이 다중회귀모델이기 때문에 독립성 가정을 포함하여 선형성, 잔차의 정규성, 잔차의 등분산성 가정을 모두 만족하는지 확인한다.

▷ 선형성

아래의 표는 각 설명변수가 기대수명과 선형관계에 있는지 확인하기 위해서 반응변수와 설명변수 사이의 상관계수를 보여준다.

features	Correlation	features	Correlation
Schooling	0.7	Status	−0.3
Adult_Mortality	−0.7	Alcohol	0.3
Income_Composition_Of_Resources	0.8	HIV_AIDS	−0.3
Diphtheria	0.5		

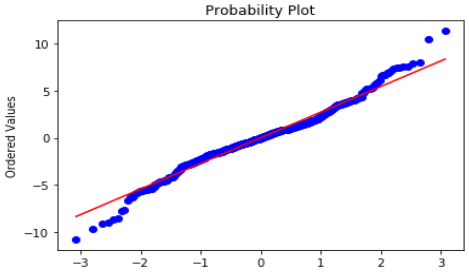
▷ 독립성

설명변수들 간의 독립성을 확인하기 위해 VIF지수를 이용한다. 아래의 표는 각 설명변수의 VIF값을 보여준다. VIF값이 5를 넘지 않는 작은 값이기 때문에 독립성 가정을 만족한다고 판단할 수 있다.

features	VIF Factor	features	VIF Factor
Schooling	2.953	Status	1.195
Adult_Mortality	1.357	Alcohol	1.194
Income_Composition_Of_Resources	2.919	HIV_AIDS	1.192
Diphtheria	1.471		

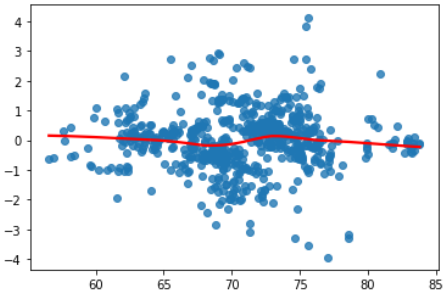
▷ 잔차의 정규성

오른쪽 그래프는 회귀모델의 잔차가 정규성을 만족하는지 확인하기 위해 그린 Q-Q Plot이다. 점이 찍힌 형태가 빨간색 직선에 근접함을 볼 수 있다. 따라서 회귀모델의 잔차가 정규성을 따른다고 볼 수 있다.



▷잔차의 등분산성

오른쪽 그래프는 회귀모델의 잔차가 등분산성을 만족하는지 확인하기 위해 그린 그래프이다. 빨간 선은 잔차의 경향성을 나타내는 선으로, 이 선이 비교적 수평에 가까우므로 잔차가 등분산성을 따른다고 볼 수 있다.



▷잔차의 독립성

모델의 잔차의 독립성 가정을 따르는지 확인하기 위해 시행한 더빈왓슨 검정 결과는 오른쪽 표와 같다. 그 결과, 유의확률이 0.05보다 크기 때문에 단순선형회귀모델이 잔차의 독립성을 만족한다는 것을 확인할 수 있다.

Durbin Watson Test	
DW	2.008
P-Value	0.5415

Ⅴ 결론

(1) 분석 결과 해석

분석 결과, 각 나라의 기대수명을 예측함에 있어 중요한 영향력을 끼친 상위 5개의 변수는 저체중인 어린이의 비율, HIV_AIDS로 인한 사망률, HRDI, 성인 사망률, 선진국 여부이다. 아시아 국가의 경우, 기대수명의 요인에 중요한 영향력을 끼친 상위 4개의 변수는 성인 사망률, HRDI, 평균 교육 기간, 선진국 여부이다. 전체 국가의 기대수명 예측과 아시아 국가의 기대수명 요인분석에서 공통적으로 큰 영향을 끼친 변수는 성인사망률, HRDI, 선진국 여부이다.

이와 같은 결과를 통해 각 나라에서 사람이 사고나 질병으로 사망하는 비율과 인적 자원으로서의 가치, 국가가 얼마나 선진화되었는지에 따라 해당 국가의 기대수명이 예측될 수 있다. 특히, HRDI 와 선진국 여부 변수가 기대수명에 큰 영향을 끼치는 것을 보아 노동력 또는 인적 자원으로서의 가치와 각 나라의 의학기술, 치안 등의 발전 정도가 중요한 요인임을 추론해 볼 수 있다.

(2) 분석 한계 및 시사점

분석에 사용한 데이터는 WHO와 UN에서 수집된 자료로 생성되었다. 데이터의 특성상 질병에 대한 정보가 많은데, WHO에 해당하는 질병들에 대한 최신 자료는 공개되지 않은 상태이다. 또한, 과거에는 AIDS 등의 질병이 사망 위험성이 높았지만 오늘날 약이 개발됨에 따라 질병의 위험성이 감소된다. 그러나 동시에 새로운 바이러스로 인해 질병이 발병하여 사망에 이른다. 이러한 질병에 대한 정보는 사망 위험성과 관련하여 예측하지 못하는 측면이 있다. 실제로 최근에 발병한 코로나19 바이러스의 위험성과 사망률을 미리 예상하지 못했다.

보도자료에 따르면 대한민국의 경우, 기대수명은 점차 증가하는 추세이지만 '건강수명'[3] 즉, 평균수명에서 질병이나 부상으로 인하여 활동하지 못한 기간을 뺀 기간은 감소하고 있는 추세이다. 뿐만 아니라, 대한민국의 기대수명과 건강수명은 소득 계층별, 지역별 격차가 뚜렷하다.[4] 이는 수

[3] [NAVER 지식백과] "건강수명" (https://terms.naver.com/entry.nhn?docId=1201082&cid=40942&categoryId=31611)

[4] 고소득자-저소득자 건강수명 11년 격차, 연합뉴스(https://www.yna.co.kr/view/AKR20200114181300017?input=1195m)

명을 분석하고 예측함에 있어, 단순히 질병 발병률이나 선진국 여부만이 아닌, 각 국가의 사회적 특성이 반영되어야 함을 보여준다.

그러나 이 분석의 목적은 개개인의 기대수명을 예측하는 것이 아닌, 특정 국가의 평균적인 기대수명을 예측하는 것이다. 따라서 구축한 모형을 이용하여 각종 질병에 대한 자료와 국가의 발전 수준으로 해당 국가의 대략적인 기대수명을 예상할 수 있다.

분석 결과, 국가 수준에 따라 국민의 기대수명이 크게 결정됨을 알 수 있다. 국가 수준이 높으면 의료 기술도 발전한 국가이기 때문에 질병으로 인한 사망 가능성이 낮고 약이 개발되어 치료받는 경우가 많기 때문이다. 즉, 국가 수준이 질병으로 인한 사망 가능성으로 직결된다고 볼 수 있다. 실제로 부유하고 선진화가 진행된 나라에서 유행하는 질병에 대한 약은 빠르게 개발되고 비교적 약소하고 선진화되지 않은 나라에서 발병하는 질병에 대한 약은 개발이 더딘 상황이다.

이 분석을 통해 국가에 따라 기대수명이 다른 가장 큰 이유는 결국 '국가 간 불균형'임을 알 수 있다. 각 국가의 기대수명을 높이기 위해서는 '교육'과 '국가의 기술력'에 집중해야 한다. 기술력의 발전은 질병에 의한 국민의 생존율을 증가시키고 이러한 기술의 발전은 교육의 발전을 통해 이룰 수 있을 것이다. 뿐만 아니라 기대수명과 건강수명을 동일하게 유지시키는 것이 모든 국가들의 과제가 될 것이다.