다중선형회귀분석을 통한 국가별 기대수명 모델링.

2015111593
경영학부
김정의

목차

제 1장 서론

1. 연구의 목적

개인의 수명에 영향을 미치는 요소는 무엇일까? 흔히 식습관, 유전적 특성, 영양적 요소등을 생각할 수 있을 것이다. 그렇다면 수준을 개인에서 국가로 넓혀 한 국가의 기대수명에 가장 영향을 미치는 요소는 무엇일까? 개인의 경우와 비슷하겠지만 좀 더 거시적으로 국가의 보건정책, 전쟁의 위험성, 지역적 특성 등등을 생각해 낼 수 있을 것이다.

현재 한 국가의 기대수명을 구하기 위해서는 각 나이별 사망률을 모두 알아야 국가의 기대별 수명을 구할 수 있다. 하지만 이러한 특성들이 국가의 기대수명에 미치는 영향을 수치로 표현하여 모델링을 한다면 좀 더 간편한 방법으로 기대수명을 예측해 볼 수 있지 않을까? 라는 생각을 하게 되었다.

그래서 **국가별 기대수명을 각 나이별 사망률을 이용하여 구하는 방식이 아니라 몇 개의 변수만을 이용하여 간편하지만 정확하게 구할 수 없을까?** 라는 주제로 연구를 진행하게 되었다.

2. 사용 데이터 설명

데이터는 Kaggle - https://www.kaggle.com/kumarajarshi/life-expectancy-who 에서 발견한 WHO의 데이터를 사용하였다. WHO 산하 'The Global Health Observatory (GHO)' 의 데이터저장소에 있는 데이터이다.

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.1
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.(
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.1
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.9
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.9

5 rows × 22 columns

그림 1 - 원 데이터의 모습

각 국가별 그리고 연도별로 기대수명, 성인사망자수, 알콜섭취량 등이 나타나있다. 이번 연구에서는 2014년의 데이터를 대상으로 진행하였다. 그 이유는 2014년 데이터도 결측값이 전혀 없는 것은 아니지만 2015년의 데이터는 결측값이 너무 많이 존재하였다. 그래서 2015년을 제외한 가장 최근인 2014년 데이터를 이용하였다. 2014년 데이터를 이용하여 모델을 설정하고 이 모델을 2015년의 데이터와 비교하

며 혹은 2014년 중에 모델에 포함시키지 않았던 데이터들을 이용하여 후에 모델을 평가하고자 한다.

3. 기대수명과 평균수명의 정의

기대수명과 평균수명의 용어의 정의를 짚고 갈 필요가 있다고 생각했다. 평균수명은 죽은 사람들의 나이를 평균을 내어 구한 값이다. 사고로 죽거나, 질병에 의해 죽은 사람들까지 모두 포함하게 된다. 기대수명은 조금 풀어서 설명하자면 현재 0세인 아이가 몇 세에 죽을 것이라고 예상되는 기대치이다. 구하는 방법은 다음과 같다. 0세의 아이의 사망률이 1%, 1세의 아이가 2% 라고 하자. 0세의 아이가 0세에 죽을 확률은 1% 이다. 0세의 아이가 2세에 죽을 확률은 (1-0.99)*(0.02)이다. 이런 식으로 각 나이별 사망률을 구한 후 기댓값을 구하여 0세의 아이의 기대수명을 구하고 이를 줄여서 기대수명이라고 부른다. 이 연구에서 사용한 것은 기대수명이다.

4. 연구의 방법

이 연구의 최종 목적은 국가의 기대수명을 잘 예측하면서 간결한 예측모델을 만드는 것이다. 그 방법은 수업주제에 맞게 회귀모형만을 사용하고자 한다. 중간보고서까지는 WHO 데이터를 이용하여 국가 전체에 대하여 다중선형회귀모델을 통하여 모델을 만들고자 한다. 기말보고서에서는 좀 더 심화하여 '선진국과 후진국 사이에 기대수명에 영향을 미치는 요소가 다른지', 다르다면 '어떻게 모델에 반영하여 모델링 할 것 인지'등등 세분화되고 더욱 구체적으로 모델링을 하고자 한다.

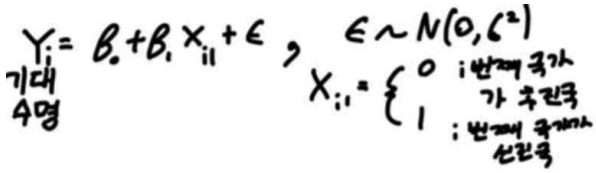
또한 다중선형회귀분석을 하고자 하는데 특히, 이 WHO 데이터는 변수들 간의 상관관계가 높아 다중공선성이 상당히 높을 것으로 예상되어 이를 잘 고려하여 변수를 선택하고 필요하다면 모델의 예측력을 더 높이기 위해 다른 데이터를 추가로 이용하여 모델을 설정하고자 한다.

데이터의 전 처리, 탐색적 자료조사에는 주로 파이썬을 사용하였다. 그리고 회귀분석, 모델링 과정에는 R을 이용하여 연구를 진행하였다.

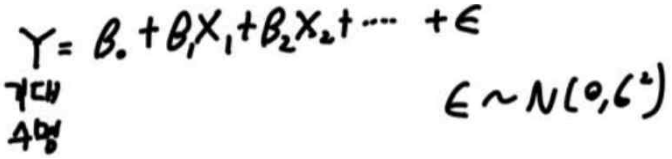
5. 예상 결과

국가별 평균수명에 미치는 영향은 개발도상국과 선진국 사이에 다르게 나타날 것이라고 예상하였다. 개발도상국에서는 당장의 식량보급률, 전쟁의 위험성 등등 당장의 생계에 연관이 된 변수들을 이용하여 모델링해야 한다고 생각했다. 이와 다르게 선

진국에서는 기본적 생존이 보장 돼있는 만큼 비만율, 흡연율, 음주율 등이 영향을 크게 미칠 수 있을 것이라고 생각하였다. 그래서 최종 모델은 이러한 형태의 모델이 나올 것이라고 예상하였다. (물론,β_0 , β_1 등 회귀계수들은 아직 알 수 없으며 훨씬 많을 것으로 예상된다.)



하지만 중간보고서까지는 국가 전체에 대하여 다중선형회귀를 통하여 모델링 하고자 한다. 식은 다음과 같다. (마찬가지로 β_0 , β_1등이 무엇이며 총 몇 개인지는 알 수 없음.)



제 2장 본론

제 1절 탐색적 자료조사 및 데이터 전처리

1. 기대수명과 변수간의 산포도

이 데이터는 총 22개의 컬럼들이 있으며 나라이름 조사연도 등을 제외하면 평균수명을 예측할 변수로 쓸 수 있는 컬럼들이 총 18개가 있다. 우선 평균수명과의 관계를 보기 위해 선택한 **18개 각각 변수들**과 **평균수명**과의 산점도를 그려보았다. 다음은, 탐색적 자료조사 결과를 간단하게 설명한 것 이다. <산점도는 부록 1 참조>

1번. 15~60세의 성인 1000명당 죽은 15~60세 성인의 수 100명 미만인 국가들을 제외하고 기대수명과 음의 상관관계를 보이고 있다.
2번. 유아 1000명당 죽은 유아의 수와 기대수명과의 관계
3번. 알콜섭취와 기대수명과의 관계, 가로축이 2이상인 부분에서 양의 상관관계가 의심된다.
4번. 1인당 국내총생산 대비 1인의 건강지출(%). 1인당 국내총생산의 값 대비 국가의 지출인지 개인의 지출인지 데이터가 명확하지 않았다.
5번. 1세 유아들의 B형간염의 예방접종 비율정도 (높을수록 접종비율이 높다.)

6번. 인구 1000명당 홍역에 감염된 사람의 수. 1000명당 홍역에 감염된 사람의 수인데 데이터 중에 1000이 넘어가는 경우가 꽤 있다.
7번. 인구전체의 평균 BMI. BMI는 연구에 꽤 도움이 될 변수라고 생각했는데 평균 BMI가 60이넘어가고 심지어 80인 국가들이 있다. 데이터에 오류가 있다고 판단했다.
8번. 인구 1000명당 5살이하 어린이들의 사망자 수. 이것 또한 1000을 넘어가는 경우가 있다. 데이터에 오류가 있을 확률이 높다고 생각했다.
9번. 1살 인구의 소아마비 예방접종 비율. 5번과 매우 유사한 분포를 보여주었다. 0프로 부근을 제외하고 양의 선형관계를 보인다.
10번. 총 GDP 대비 정부의 의료, 보건에 대한 지출의 비율. 양의 선형관계가 있다고 판단하였다.
11번. 1살 인구의 디프테리아 예방접종비율. 5,9,11은 매우 유사한 분포를 보인다.(같은 그래프 아님). 예방접종을 잘 시행하는 국가들은 위험한 질병들에 대하여 모두 시행할 확률이 높기 때문에 그러한 것 이라고 생각하였다.
12번. 0~4살 인구 1000명당 에이즈로 인해 죽은 0~4살 인구의 수.
13번. 국가별 1인당 GDP. 어느 정도 기대수명과 연관이 있는 것 같지만 1인당 GDP 가 지나치게 커질 경우 기대 수명에 큰 영향을 미치지는 못하는 것처럼 보였다.
14번. 국가별 인구. 인구가 너무 큰 국가들을 제외하고 4천만 이하에서 산포도를 그려봤지만 어떠한 패턴이 보이진 않았다.
15번. 1~19세 인구중 영양실조에 걸린 인구의 비율.
16번. 5~9세 인구중 영양실조에 걸린 인구의 비율.
17번. 인간개발지수. 인간개발지수를 산출할 때 기대수명을 직접적으로 이용한다고 한다. 이 변수를 포함하면 a의 상태를 a를 통해 예측하는 것이 되므로 제외하기로 하였다.
18번. 국가별 한 일생동안 교육을 받는 평균연도의 값. 성인 전의 교육기간만을 얘기하는 것이 아니다. 한 눈에 봐도 선형적인 관계가 나왔다.

2.데이터 전처리

국가들 중에서 남수단과 수단은 결측치를 많이 가지고 있는 국가들이여서 제외하기로 하였다. 또한 18번 변수의 경우 18번 변수의 값들 안에서 10개의 나라가 값을 가지고 있지 않아 이 10개의 나라를 제외했다.
(Côte d'Ivoire, Czechia , Democratic People's Republic of Korea , Democratic Republic of the Congo , Republic of Korea , Republic of Moldova , Somalia, 영국, United Republic of Tanzania, United States of America)
영국, 한국, 미국 등 주요나라가 포함되어 있었지만 18번의 변수가 산점도상으로 보았을 때 꽤 영향력 있는 변수로 생각되어 어림잡아 값을 넣는 것보다는 차라리

이들 나라를 제외하고 모델링을 한 후에 모델을 검증할 때 이 데이터들을 직접 찾아 모델 검증에 사용하고자 했다.

또한 열중에서는 13번열 '1인당 GDP' 는 지우지 않고 14번열 population 열과 파란색 열 5번열(B형간염)을 추가적으로 제거했다.

13번열(1인당 GDP) 같은 경우 에초에 연구를 시작할 때부터 매우 연관이 있을 것이라고 생각했다. 하지만 데이터 상에서 GDP는 28개의 결측치를 가지고 있었다.

14번열 Population은 42개의 결측치가 있었다. 183개의 나라에서 이미 12개 가량이 빠진 상황에서 총 42개의 관측치가 추가로 빠지게 된다면 너무 많은 관측치가 빠지게 될 것 같다는 생각이 들었다. 그래서 산포도상에서 기대수명과 더욱 연관이 있어 보이며 결측치가 더 적은 13번열 'GDP' 는 지우지 않고 14번열은 지워야겠다고 판단하였다.

5번의 경우 결측치가 10개 존재하였다. 비교적 적지만 11번 디프티레아와 상관계수가 0.877,9번 소아마비 0.691로 상당히 높아 제거한다고 해도 11번, 9번 변수를 이용하여 대체할 수 있을 것 이라고 생각했다.

결론적으로, 데이터에 오류가 있어보이는 빨간색 열 들과 결측값이 있는 파란색 열들은 제외 하였다. 그 결과 빨간색 열과 파란색 열을 제외한 총 11개의 변수들과 또한 결측치가 있는 국가들 을 제외하고 남은 152개의 국가들을 이용해서 모델을 만들기로 하였다.

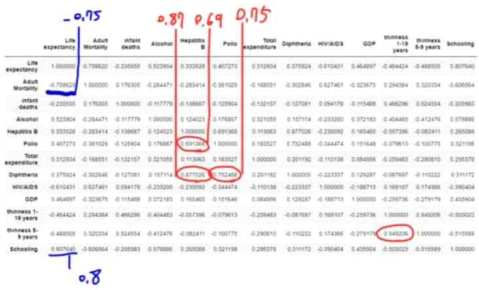
제 2절 상관분석 및 다중공선성

1. 상관분석

연구를 시작하기 전부터 변수들 간의 상관관계가 매우 커서 다중공선성의 위험을 가장 걱정했다. 11개 변수들 중 알코올섭취량 정도를 제외하면 변수들 사이의 상관관계가 클 것으로 예상됐다. 예를 들어, 1인당 GDP가 높은 나라는 생계외에 쓸 수 있는 돈이 많으므로 교육에 쓸 수 있는 돈도 많으므로 GDP와 교육연도는 양의 상관관계를 가질 것이다. 이 외에도 5번이나 9번변수 (1세 인구의 B형간염 면역정도, 1세의 소아마비의 면역정도) 는 사실상 1세의 건강에 신경쓰는 나라라면 두 변수 모두 높게 나올 것이기 때문에 매우 상관성이 높을 것으로 예상했다. 다음은 변수들 을 삭제하고 남은 변수들 과 기대수명의 피어슨상관계수를 나타낸 행렬이다.

그림 2 - 기대수명 및 변수들 사이의 상관행렬

	Life expectancy	Adult Mortality	infant deaths	Alcohol	Hepatitis B	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	thinness 1-19 years	thinness 5-9 years	Schooling
Life expectancy	1.000000	-0.758620	-0.235555	0.523904	0.333528	0.407273	0.312934	0.375924	-0.610431	0.464897	-0.464424	-0.488505	0.807640
Adult Mortality	-0.758620	1.000000	0.176305	-0.284471	-0.283414	-0.381029	-0.168551	-0.302846	0.627461	-0.323675	0.294384	0.320334	-0.606564
infant deaths	-0.235555	0.176305	1.000000	-0.117779	-0.138687	-0.125904	-0.132157	-0.127081	0.094178	-0.115488	0.466296	0.524554	-0.205983
Alcohol	0.523904	-0.284471	-0.117779	1.000000	0.124023	0.176867	0.321055	0.157114	-0.233200	0.372183	-0.404483	-0.412476	0.578886
Hepatitis B	0.333528	-0.283414	-0.138687	0.124023	1.000000	0.691368	0.113063	0.877026	-0.230092	0.165465	-0.057396	-0.082411	0.265088
Polio	0.407273	-0.381029	-0.125904	0.176867	0.691368	1.000000	0.183527	0.752488	-0.344474	0.151648	-0.079613	-0.100775	0.321198
Total expenditure	0.312934	-0.168551	-0.132157	0.321055	0.113063	0.183527	1.000000	0.201192	-0.110138	0.084956	-0.259483	-0.290810	0.295378
Diphtheria	0.375924	-0.302846	-0.127081	0.157114	0.877026	0.752488	0.201192	1.000000	-0.223337	0.129287	-0.087697	-0.110222	0.311172
HIV/AIDS	-0.610431	0.627461	0.094178	-0.233200	-0.230092	-0.344474	-0.110138	-0.223337	1.000000	-0.188713	0.169107	0.174366	-0.390404
GDP	0.464897	-0.323675	-0.115488	0.372183	0.165465	0.151648	0.084956	0.129287	-0.188713	1.000000	-0.259736	-0.279179	0.435904
thinness 1-19 years	-0.464424	0.294384	0.466296	-0.404483	-0.057396	-0.079613	-0.259483	-0.087697	0.169107	-0.259736	1.000000	0.949206	-0.503023
thinness 5-9 years	-0.488505	0.320334	0.524554	-0.412476	-0.082411	-0.100775	-0.290810	-0.110222	0.174366	-0.279179	0.949206	1.000000	-0.515589
Schooling	0.807640	-0.606564	-0.205983	0.578886	0.265088	0.321198	0.295378	0.311172	-0.390404	0.435904	-0.503023	-0.515589	1.000000



몇 가지만 살펴보면,

1. Polio, 디프테리아,B형간염사이에 상관계수가(0.87,0.69,0.75)로 매우 높은 상관관계를 가짐.
2. 1~19세의 영양실조와 5~9세의 영양실조도 0.949로 매우 높은 상관성을 가짐.
3. [adult mortality와 schooling]의 기대수명과의 관계가 각각 -0.75, 0.80으로 매우 높은 상관관계를 가졌다.

2. 다중공선성 확인

변수들 사이의 상관관계가 꽤 높은 결과를 가지므로 변수들 간의 다중공선성이 발생할 확률이 충분히 있다고 생각하였다. 그래서 VIF를 이용하여 다중공선성이 있는지부터 판단하기로 하였다. 11개의 변수만을 이용하여 모델을 만들 것이므로 11개의 변수에 대하여 값을 구하였다. R로 VIF값을 구한 결과 다음과 같았다.

```
> vif(lm_14)
     Adult.Mortality          infant.deaths                Alcohol                  Polio
            2.402120              1.497924              1.716239              2.686736
Total.expenditure               Diphtheria               HIV.AIDS                   GDP
            1.194627              2.489777              1.818726              1.299266
thinness..1.19.years thinness.5.9.years             Schooling
            9.256213             10.518734              2.689624
```

그림3 - 11개 변수들의 VIF

일반적으로 10을 넘어가면 다중공선성이 심각하다고 판단을 하고 4를 기준으로 삼기도 한다고 한다. 처음 생각과는 다르게 전체적으로 값이 높게 나오지 않아 다중공선성을 굳이 고려하지 않아도 된다고 판단하였다. 하지만 영양실조와 관련된 2변수에서 각각 10.212, 11.437로 다중공선성이 확인되었고 두 변수 중에 하나만을 사용하기로 하였다. 그리하여 5~9세의 인구를 포함하고 있는 1~19세의 영양실조

인구 변수만을 우선 모델에 넣기로 하였다.

제 3절 변수 선택 및 모델적합

변수선택에서는 후진선택법을 이용해보기로 하였다. 우선 모든 변수를 넣어서 모델을 만들어보고 후에 설명력이 가장 떨어지는 설명변수들을 하나씩 제거하였다.
모든 변수를 넣어 나온 처음 모델은 다음과 같았다.

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.6524  -1.6518   0.0681   2.1071   8.1668

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.751e+01  2.650e+00  21.698  < 2e-16 ***
Adult.Mortality     -2.784e-02  4.154e-03  -6.703 4.54e-10 ***
infant.deaths        1.068e-03  3.621e-03   0.295 0.768461
Alcohol              1.478e-01  9.244e-02   1.598 0.112183
Polio               -9.577e-03  2.279e-02  -0.420 0.674986
Total.expenditure    2.363e-01  1.190e-01   1.985 0.049090 *
Diphtheria           3.184e-02  2.120e-02   1.502 0.135271
HIV.AIDS            -1.038e+00  2.666e-01  -3.896 0.000151 ***
GDP                  4.435e-05  1.779e-05   2.493 0.013838 *
thinness..1.19.years -1.549e-01  9.162e-02  -1.690 0.093154 .
Schooling            1.182e+00  1.674e-01   7.060 6.94e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.577 on 141 degrees of freedom
Multiple R-squared:  0.8428,    Adjusted R-squared:  0.8316
F-statistic: 75.59 on 10 and 141 DF,  p-value: < 2.2e-16
```

그림 4 - 시작상태 모형의 summary 결과.

각 설명변수들에 대응하는 p-value 값중에 가장 높은 값이 나온 infant.deaths(0.768) 부터 제거하여 p-value 값이 0.05보다 높게 나온 설명변수가 없을 때까지 제거해 보았다.

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + Total.expenditure +
    HIV.AIDS + GDP + Schooling, data = df14)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3465  -1.7217  -0.0748   2.1003   9.5166

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.521e+01  2.084e+00  26.494  < 2e-16 ***
Adult.Mortality   -2.688e-02  4.139e-03  -6.493 1.23e-09 ***
Total.expenditure  3.027e-01  1.170e-01   2.588  0.01062 *
HIV.AIDS          -1.068e+00  2.665e-01  -4.006 9.80e-05 ***
GDP                5.077e-05  1.773e-05   2.863  0.00482 **
Schooling          1.444e+00  1.422e-01  10.158  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.627 on 146 degrees of freedom
Multiple R-squared:  0.8326,    Adjusted R-squared:  0.8268
F-statistic: 145.2 on 5 and 146 DF,  p-value: < 2.2e-16
```

그림 5 - 후진제거법을 통한 모델① 의 summary 결과.

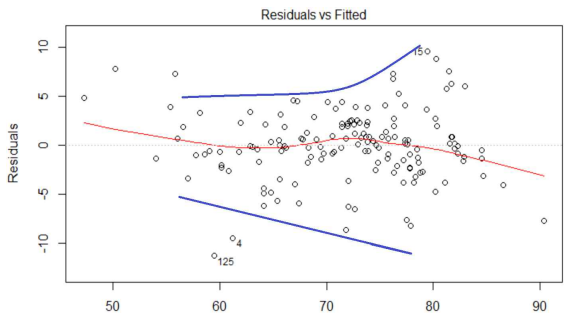


그림 6 - residuals vs fitted values 그래프

결과적으로 1번 Adult.mortality(성인사망자수), 10번 Total.expenditure(정부의 지출비율, 12번 AIDS 감염자수 , 13번 1인당 GDP , 18번 교육연수 만이 남게 되었다. (이것을 모델 ① 이라 하겠다.)
수정결정계수 값도 0.831에서 0.826으로 크게 떨어지지 않았으며 0.8 이상이므로 설명력이 나쁘지 않다고 판단하였다. 또한 각 설명변수들도 유의하게 나왔다. 하지만 fitting 된 값과 잔차들의 산포도가 이분산성을 보인다고 생각하였다.
(그림 6 의 파란선)

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + HIV.AIDS + GDP +
    Schooling, data = df14)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3388  -2.1934  -0.0803   1.9991   9.4085

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.561e+01  2.118e+00  26.261  < 2e-16 ***
Adult.Mortality -2.616e-02  4.209e-03  -6.215 5.03e-09 ***
HIV.AIDS        -1.074e+00  2.716e-01  -3.954 0.000119 ***
GDP              4.860e-05  1.805e-05   2.692 0.007920 **
Schooling        1.549e+00  1.389e-01  11.148  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.697 on 147 degrees of freedom
Multiple R-squared:  0.8249,    Adjusted R-squared:  0.8201
F-statistic: 173.1 on 4 and 147 DF,  p-value: < 2.2e-16
```

그림 7 - 그림 5 에서 total.expenditure 를 제거한 모델②의 summary결과

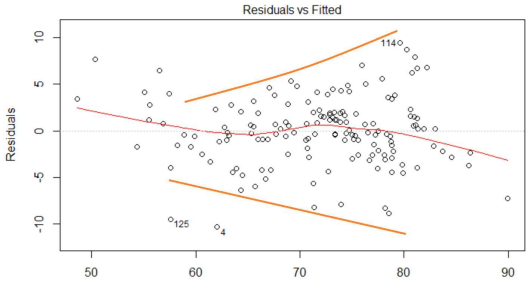


그림 8 - 그림7의 resdiduals vs fitted values 그래프

추가적으로 유의한 설명변수들중 그림 5에서 가장 p-value가 높았던 10번 변수 Total.expenditure(0.01)을 제거한 모델을 구해 보았다(그림 7). 마찬가지로 모든 계수들이 유의미하였다. 또한 결정계수의 값도 0.82로 충분히 높고 모델 ① 과 0.06정도의 차이밖에 나지 않았다. 하지만 전 모델처럼 (그림 8)에서 다음과 같이 이분산성이 발견되었다. 여기서 추가적으로 변수를 제거할 경우 결정계수의 값이 크게 내려가게 되어 더 이상 변수를 빼서는 안 된다고 판단을 하였다. 그래서 이분산성을 제거하기 위해 변수변환을 하기로 결정하였다. 그중에서 12번 그래프였던 HIV.AIDS 변수가 지나치게 비선형적이기 때문에 모델을 적합하는데 있어 문제가 되지 않을까 생각했다 (그림9). 그래서 HIV.AIDS 변수를 산포도와 비슷한 함수인 1/x 로 표현하기 위해 HIV.AIDS 변수를 1/HIV.AIDS 값으로 변환하여 회귀모델을 돌려보았다. (1/HIV.AIDS = HIV_1로 지칭함)

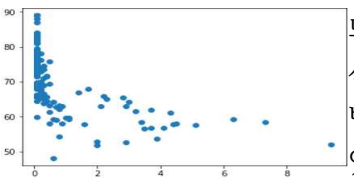


그림 9 - <부록1> 12번 그래프

모델③: ①에서 12번 HIV.AIDS을 역수로 변환한 모델 (10번 변수를 포함한 모델)

모델④: ②에서 12번 HIV.AIDS을 역수로 변환한 모델 (10번 변수를 제외한 모델)

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + Total.expenditure +
    HIV_1 + GDP + Schooling, data = df14)

Residuals:
     Min       1Q   Median       3Q      Max
-9.4323 -1.8736 -0.1241  1.6818  9.6356

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.474e+01  2.039e+00  26.850  < 2e-16 ***
Adult.Mortality  -2.739e-02  3.787e-03  -7.234 2.45e-11 ***
Total.expenditure  3.185e-01  1.142e-01   2.790  0.00597 **
HIV_1             5.424e-01  1.104e-01   4.915 2.36e-06 ***
GDP               4.816e-05  1.729e-05   2.786  0.00605 **
Schooling         1.134e+00  1.538e-01   7.376 1.13e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.54 on 146 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.8351
F-statistic: 153.9 on 5 and 146 DF,  p-value: < 2.2e-16
```

그림 10 -③의 summary 결과

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + HIV_1 + GDP +
    Schooling, data = df14)

Residuals:
     Min       1Q   Median       3Q      Max
-9.3316 -1.9915  0.1176  1.9205  9.4731

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.518e+01  2.079e+00  26.542  < 2e-16 ***
Adult.Mortality -2.680e-02  3.867e-03  -6.931 1.22e-10 ***
HIV_1            5.361e-01  1.129e-01   4.750 4.80e-06 ***
GDP              4.587e-05  1.767e-05   2.597   0.0104 *
Schooling        1.248e+00  1.517e-01   8.231 9.19e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.621 on 147 degrees of freedom
Multiple R-squared:  0.832,     Adjusted R-squared:  0.8275
F-statistic:   182 on 4 and 147 DF,  p-value: < 2.2e-16
```

그림 11- ④의 summary 결과

변수변환 결과 수정 결정계수의 값도 두 모델에서 모두 상승하였으며 설명변수들도 유의미 한 것으로 나타났다. 또한 두 경우에서 모두 이분산성이 제거되었으며 분포도 고르게 분포하는 모습을 보였다(그림12, 그림13).

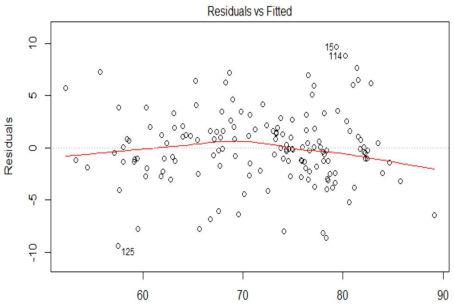


그림 12 - ③의 Residuals vs fitted values 그래프

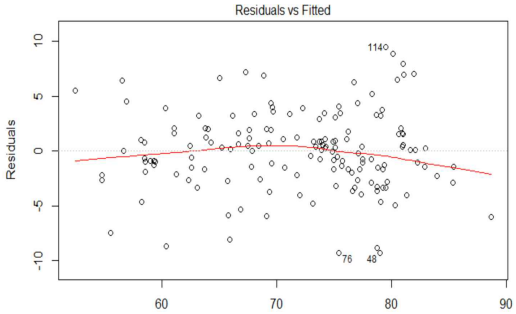


그림 13 - ④의 Residuals vs fitted values 그래프

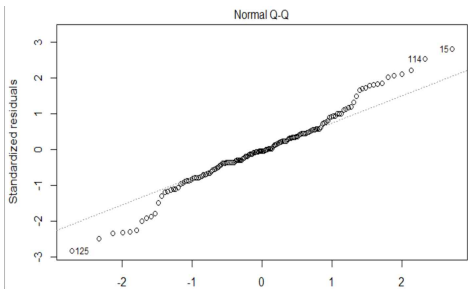


그림 14 - ③의 Q-Q plot

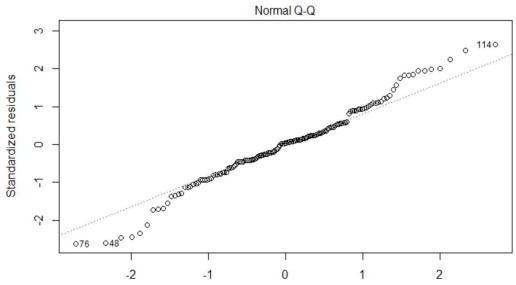


그림 15 - ④의 Q-Q plot

```
        Shapiro-Wilk normality test

data:  residual_with_expenditure
W = 0.98005, p-value = 0.02643
```

그림 16 - ③의 정규성 검정

```
        Shapiro-Wilk normality test

data:  residual_none_expenditure
W = 0.98496, p-value = 0.09754
```

그림 17 - ④의 정규성 검정

그러나 그림 14, 그림 15의 Q-Q plot을 그려본 결과 양쪽 꼬리부분에서 기준선과 거리가 꽤 있는 것으로 보였다. 정규성을 검증하기 위해 추가로 Shapiro-wilk 테스트를 진행하였다. 결과적으로 ④는 유의수준 0.05에서 0.09754>0.05로 '정규분포를 따른다고 볼 수 있었지만 ③의 경우 0.02643 < 0.05 로 정규분포를 따른다고 말할 수 없었다. 그리하여 모델③ 대신 모델④를 택하는 것이 적합하다고 판단했다.

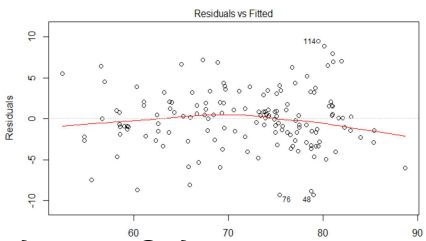


그림 13 - ④의 residuals vs fitted values 그래프

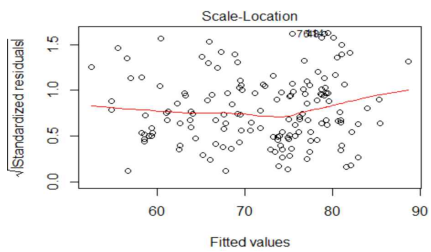


그림 18 - ④의 scale -location plot

또한 그림 13, 18을 통해 잔차들이 전체적으로 고르게 분포한 것을 보아 잔차의 등분산성을 확인할 수 있었다.

결론적으로 수정 결정계수 0.83 으로 설명력이 높으며 회귀분석의 기본가정을 만

족하는 ④모델을 택하기로 하였다.
최종모델 = ④
④ -> 기대수명 = 55.18 -0.027*(15~60세 성인사망자수)+ 0.54*1/(AIDS감염자수)+0.000046*(1인당GDP)+1.25*(평균교육연수)+ ϵ

3장 결론

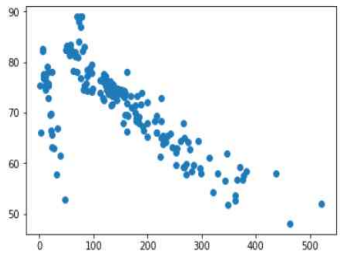
국가별 기대수명을 최종적으로 변수 4개를 통하여 표현하는 다중선형회귀모형을 만들 수 있었다. 모델의 설명력이 높고 기본가정도 잘 만족하여 만족스러웠다. 원래 모델의 평가까지도 중간보고서의 범위에 넣으려 했지만 모델의 평가까지 진행하기에는 분량이 지나치게 길어질 것 이라 생각하였다. 기말보고서에서는 개발도상국과 선진국으로 분리하여 구하는 등 좀 더 정교한 모델을 만들 것이다. 모델의 평가는 정교해진 모델까지 포함하여 같이 진행을 하기로 하였다.

또한 이 모델의 경우 WHO의 데이터만을 사용하여 모델을 설정하였는데 이 변수들이 국가별 기대수명을 모델링 하는데 최선의 변수인지 그리고 더 좋은 변수는 없는지 등을 추가적으로 탐색하고 필요시 WHO데이터 이외의 데이터를 추가적으로 사용하여 모델을 만들고자 한다.

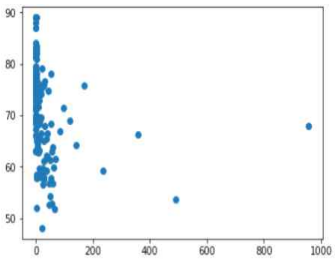
<부록>

1.기대수명과 변수들과의 scatter plot

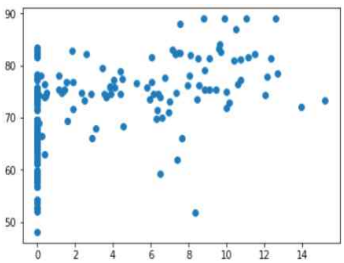
1



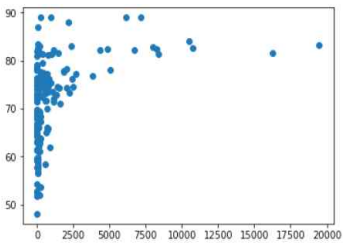
2



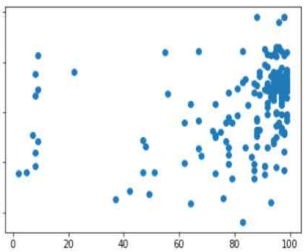
3



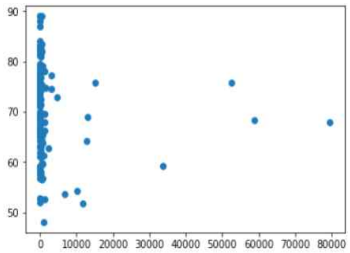
4



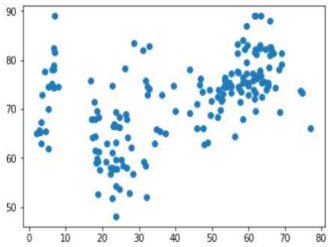
5



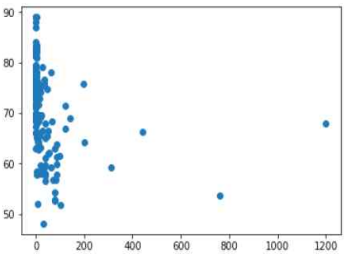
6



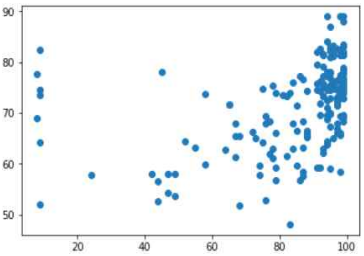
7



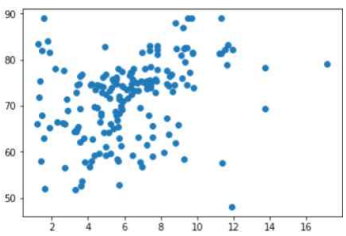
8



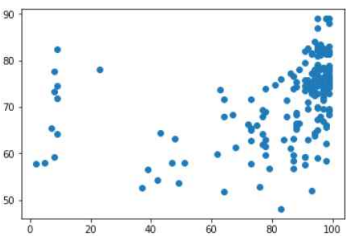
9



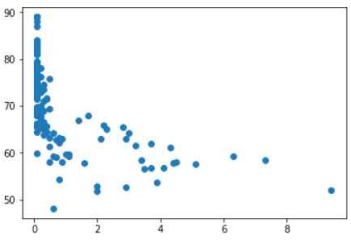
10



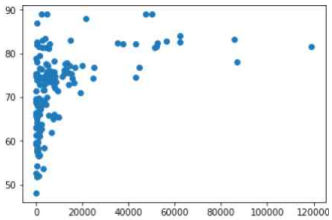
11



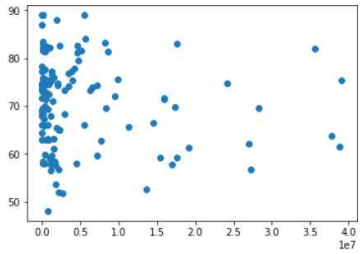
12



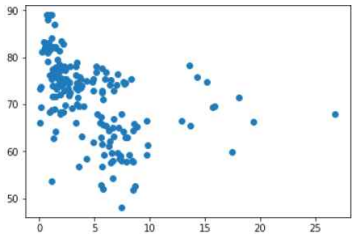
13



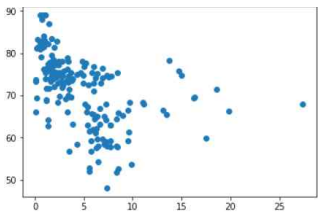
14



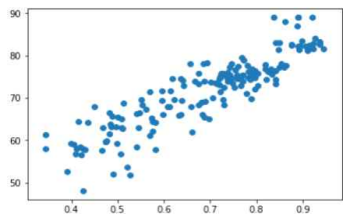
15



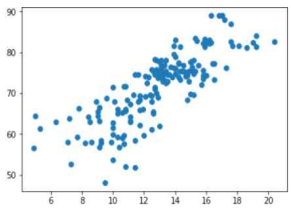
16



17



18



2. 사용 코드

2-1 전처리과정 & 상관계수 구하는 부분 (파이썬 사용)

```
#데이터 부르기
df0=pd.read_csv('life.csv')

#2014년인 데이터만 부르기
df_2014=df0.loc[df0['Year']==2014]

# 기대수명과 변수들과의 산포도를 그리는 것은 matplotlib을 이용하여 18개 각각 구함.
# 예시
plt.scatter(df_2014['Status'],df_2014['Life expectancy '])

#21번째 열에 없는 데이터 10개 짜르기
df_14_1=df_2014.loc[df_14[df_14.columns[11]].isnull()==False]

# 데이터가 적었던 남수단 수단 짜르기
df_140=df_14_1.drop([2410,2458]) #Lesotho Angola

# 상관행렬 만들기
df_140[['Life expectancy ', 'Adult Mortality',
        'Alcohol', 'Hepatitis B',
        'Polio', 'Total expenditure',
       'Diphtheria ', ' HIV/AIDS', 'GDP',
       ' thinness  1-19 years', ' thinness 5-9 years',
```

```
'Schooling']].corr()



# 필요없는 column 들 자르기.
del df_140['Country']
del df_140['Status']
del df_140['percentage expenditure']
del df_140['Hepatitis B']
del df_140['Measles ']
del df_140[' BMI ']
del df_140['under-five deaths ']
del df_140[ 'Polio']
del df_140['Population']
del df_140[' thinness 5-9 years']
del df_140['Income composition of resources']

# csv 파일로 변형
df_140.to_csv(life__14.csv,index=False)
```

2-2 변수선택 및 회귀분석 (R 이용)

```
# 데이터부르기
df14 <- read.csv('C:/Users/Jung Eui/Desktop/life-expectancy-who/life__14.csv',header=T)
str(df14)
head(df14)

# 다중공선성 체크
library(car)

#lm_14 은 모두 넣은 모델
lm_14 <- lm(Life.expectancy~.,data=df14)
vif(lm_14)

# lm1 최초 시작모델
lm1 <-
lm(Life.expectancy~Adult.Mortality+infant.deaths+GDP+Polio+Total.expenditure+Diphtheria+HIV.
AIDS+thinness..1.19.years+Schooling,data=df14)
summary(lm1)
lm_2 <-
lm(Life.expectancy~Adult.Mortality+infant.deaths+Alcohol+Polio+Total.expenditure+Diphtheria+H
IV.AIDS+GDP+Schooling,data=df14)
summary(lm_2)
lm_3 <-
lm(Life.expectancy~Adult.Mortality+Alcohol+Polio+Total.expenditure+Diphtheria+HIV.AIDS+GDP
+Schooling,data=df14)
```

```
summary(lm_3)
lm_4 <- lm(Life.expectancy~Adult.Mortality+Alcohol+Total.expenditure+Diphtheria+HIV.AIDS+GDP+Schooling,data=df14)
summary(lm_4)
lm_5 <- lm(Life.expectancy~Adult.Mortality+Alcohol+Total.expenditure+HIV.AIDS+GDP+Schooling,data=df14)
summary(lm_5)

# lm_6 = 후진적제거법의 결과로 나온 모델 = 1번모델
lm_6 <- lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+GDP+Schooling,data=df14)
summary(lm_6) # 토탈포함
plot(lm_6)

# lm_7 = 추가로 total.expenditure 를 뺀 모델 = 2번 모델
lm_7 <- lm(Life.expectancy~Adult.Mortality+HIV.AIDS+GDP+Schooling,data=df14)
summary(lm_7)
plot(lm_7)

#변수변환
HIV_1 <- 1/df14$HIV.AIDS

# 변수변환한 모델 , 보고서에서 모델 3
lm_8 <- lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV_1+GDP+Schooling,data=df14)
summary(lm_8)
plot(lm_8)

#변수변환한 모델 , 보고서에서 모델 4
lm_7 <- lm(Life.expectancy~Adult.Mortality+HIV_1+GDP+Schooling,data=df14)
summary(lm_7) #최종모델 변수변환한
plot(lm_7)

# 모델 3의 shapiro.test
residual_with_expenditure <- lm_8$residuals
shapiro.test(residual_with_expenditure)

# 모델 4의 shapriro.test
residual_none_expenditure <- lm_7$residuals
shapiro.test(residual_none_expenditure)
```