다중회귀분석을 통한 국가별 기대수명 예측모형

2반 12팀
2015111593
경영학부
김정의

목차

제1장 서론----------3
제 1절 중간과제까지 내용정리 및 기말 과제의 연구방향--3
제 2절 사용데이터 설명----------3
1.데이터의 설명변수들(columns)----------3
2.데이터의 국가들 (rows)----------4
가. 국가수준분류 기준----------4
나. 사용데이터국가들 rows----------5
제 2장 본론----------6
제 1절 변수선택----------6
가. 변수선택 기준----------6
나. 산점도를 통한 변수 선택----------7
다. 상관관계 및 다중공선성 확인----------9
제 2절 Model building----------9
가. GDP변수와 Schooling 변수----------9
나. 질병변수----------11
다. 생활습관----------12
라. 회귀모형의 가정성립 확인----------13
제 3절 변수 중요도 및 모델 평가----------14
가. 변수 중요도 ----------14
나. 모델 평가----------15
제 3장 결론----------16
결론 및 해석----------16
부록----------18

서론

제 1절 중간과제까지 내용 정리 및 기말과제에서의 연구 방향

중간프로젝트까지 WHO의 기대수명 및 보건과 관련한 데이터들을 이용하여 국가별 기대수명을 다중선형회귀모형으로 예측해보았다.

기말과제에서는 우선 WHO 데이터의 부정확하다고 생각했던 변수들을 더 나은 데이터들을 이용하여 교체하였고 변수로 사용할 수 있을 법한 데이터들을 추가적으로 구하였다. 또한 기말과제에서는 모델의 향상을 위하여 국가의 수준을 고려하여 모델을 더 정교하게 만들어 보고자 한다. 방법은 모델을 선진국, 후진국, 그리고 중진국 (신흥 공업국이라고도 불리는) 이렇게 3가지로 나누어 모델링을 하고자 하는 방안을 생각해보았다. 국가의 수준에 따라 기대수명에 영향을 미치는 요소가 다를 수 있다고 생각했기 때문이다. 예를 들어, 고혈압은 후진국에서는 크게 중요한 변수가 아니지만 선진국끼리 비교하였을 때는 고혈압이 중요한 변수 일 수도 있다. 또한 만들어진 모델을 바탕으로 변수들의 중요성을 판단해보고 실제 기대수명과 모델을 이용해 구해진 기대수명과 비교해보고자 한다.

제 2절 사용데이터 설명

1. 사용데이터, 설명변수들 (columns)

중간과제에서는 Kaggle에서 발견한 WHO의 데이터들을 정리하여 올린 데이터들을 사용했었다. 이 데이터들의 대부분을 모델에 적합할 변수의 후보로써 그대로 사용을 할 것이다. 하지만 이 데이터에는 결측치나 오류로 보이는 변수들이 많이 있었다. 결측치들이 상대적으로 적었던 2014년의 데이터를 사용했기 때문에 추가적으로 구한 데이터들도 2014년을 기준으로 구해졌다는 것을 조심해주기 바란다. 데이터의 신뢰도를 걱정하고 있었는데 추가적으로 사용할 데이터를 탐색하는 과정에서 기대수명의 경우 'World Bank'의 자료를 그리고 1인당 GDP의 경우 우리나라 통계청에서 제공하는 국가별 GDP와 인구 자료를 발견 할 수 있었다. WHO의 자료와 새로 구한 데이터의 자료를 비교해서 봤을 때 큰 차이가 있었던 것은 아니었다. 하지만 데이터의 신뢰성을 위하여 새로 구한 데이터들로 바꾸는 것이 좋다고 생각하였다.

또한 중간과제에서는 데이터의 정확성 때문에 제거하였던 BMI와 국가별 알코올 섭취량들을 대체하기 위해 '국가별 국민들의 비만도 데이터' 와 '1인당 순수 알코올 섭취량' 들을 추가로 구할 수 있었다. BMI 데이터의 경우 BMI 지수가 30이 넘는 인구의 비율, 1인당 순수 알코올 섭취량의 경우 술 소비량에서 순수 알코올의 양을 Liter로 표시한 데이터이다. 이를 통하여 비만도와 알코올 섭취량을 다중선형회귀모델에 적합시킬지 고려해 볼 것이다.

마지막으로 Kaggle에서 구한 'Country health Indicator' 자료를 사용하였다. 이 데이터에는 최근의 코로나와 관련된 데이터들부터 각 질병들의 DALY 값, 야채,해산

물,견과류 등등 각 식품의 생산량, 수출량 등등이 나와 있었다. 질병들의 DALY란 장애보정손실수명으로써 질병, 장애, 또는 조기사망 등으로 건강하게 살지 못한 기간이나 손실된 기간을 10만명당 연단위로 표현한 것이다. 쉽게 생각하여 어떤 질병의 DALY 값이 10이라면 그 국가의 10만명의 인구가 그 질병에 의해 질병을 겪어 고통을 받은 기간이 10년이라고 생각하면 된다 **(각주 설명 참조)**[1]. 아쉬웠던 점은 국가들의 식습관 등을 반영하고자 회귀모델에 적합하고 싶었지만 자료의 데이터가 야채의 섭취량이 아니고 야채의 생산량, 수출량 등을 합한 값으로 표현한 데이터였다. 예를 들어 노르웨이에서 고등어를 많이 생산한다고 하여 노르웨이사람들이 고등어를 많이 섭취한다고 말할 수 가 없었으므로 이 데이터들을 사용할 수 없었다. 그래서 이 데이터에서는 질병의 DALY값들을 위주로 가져왔으며 추가적으로 담배가 사망에 영향을 준 정도에 관한 데이터를 가져오게 되었다.
자료의 출처는 **부록<1>** 을 참고하길 바란다.

2. 사용 데이터, 국가들 (Rows)

가. 국가 수준분류 기준

선진국 ,중진국 ,후진국을 분류하는 기준은 기관마다 다르다. 그래서 세계기구들의 기준을 고려하대 연구의 목적에 맞추어 자체적 기준을 세우게 되었다.

선진국: 1인당 GDP가 매우 높은 부국으로써 기대수명 증가를 위해 더 이상의 부의 축적의 영향보다는 다른 변수들이 영향을 미칠 것으로 예상되는 국가들. GDP 30000$ 이상의 국가들.(바하마스, 쿠웨이트, 카타르 ,UAE 제외)
중진국 : 선진국과, 후진국을 제외한 나라들.
후진국: 1인당 GDP가 매우 낮아 기초적 보건환경, 식량배급조차 어려움을 겪고 있는 나라. GDP 2000$ 이하의 국가들.

이러한 방법으로 국가를 나눈 이유는 선진국과 중진국으로 나눈 이유와 중진국과 후진국으로 나눈 이유 두 가지로 설명하겠다.
첫째, 80세 이상의 국가들의 경우 기대수명과 설명변수와의 산점도를 그려 관계를 살펴보아도 **그림<1>** 의 산점도와 같이 대부분의 변수들에서 패턴이 전혀 보이지 않았다. 전체적으로 선형성을 뛰거나 관계가 있어 보이는 변수들을 이용해 다중선형회귀모형을 만든다면 전체적으로 봤을 때 설명력이 뛰어난 모델을 만들 수 있겠지만 80세 이상에서는 잘 적합하지 않을 것을 걱정하였다. 그래서 수준을 80세 이하와 80세 이상으로 구분하여 80세 이상에서 뚜렷한 특징을 보여주는 변수를 적용한다면 더 좋은 모델을 만들 수 있을 것이라고 생각하였다. 그리고 이 80세 이상의 국가들은 모두 gdp가 30000불 이상이었다.

---

1) DALY산출 방식: https://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/

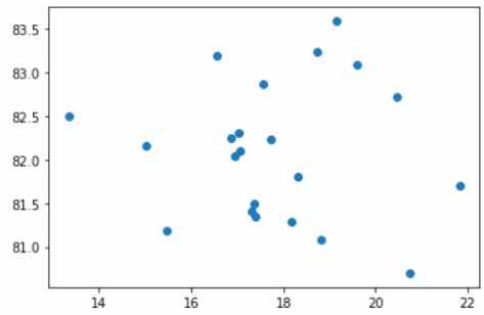


그림 <1> 80세 이상 국가들의 기대수명과 cancer변수의 산점도

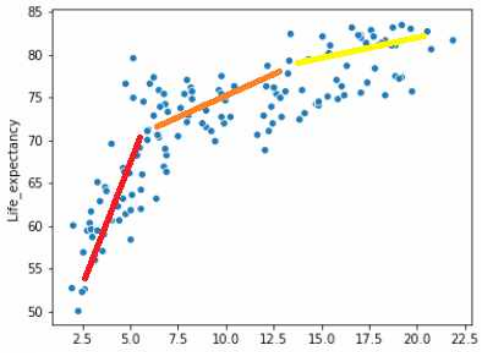


그림 <2> 기대수명과 cancer변수와의 산점도

그중에서 석유에 의해 1인당 GDP가 높게 나온 나라들 바하마스, 쿠웨이트, 카타르, UAE를 선진국에서 제외했다. 이 나라들은 오일머니에 의하여 다른 지표의 수준은 중진국 수준이지만 1인당 GDP가 비정상적으로 높았다. 이 나라들은 일반적인 경우라고 볼 수 없기 때문에 선진국에서 제외하고 중진국에 포함시켰다.

두 번째, 중진국과 후진국을 구별하고자 했던 이유는 후진국의 경우 기초적인 식량, 보건시설 조차 되어있지 않은 나라들로써 이 나라들에서는 보건지수와 관련된 것들이 다른 나라들에 비하여 크게 영향을 미쳤다. 일반적으로 IMF나 국제기구의 기준에 따르면 1300~1500$를 기준으로 후진국이 분류되지만 그렇게 되면 지나치게 후진국으로 분류되는 국가가 적었다. 2000$로 상한을 둔 것은 1인당 GDP가 2039$인 인도를 후진국으로 분류해서는 안 된다고 생각했기 때문이다. 주관적 의견이 들어간 결정이지만 대부분의 국제기구들도 일반적으로 인도를 중진국에 포함시키기 때문이다.

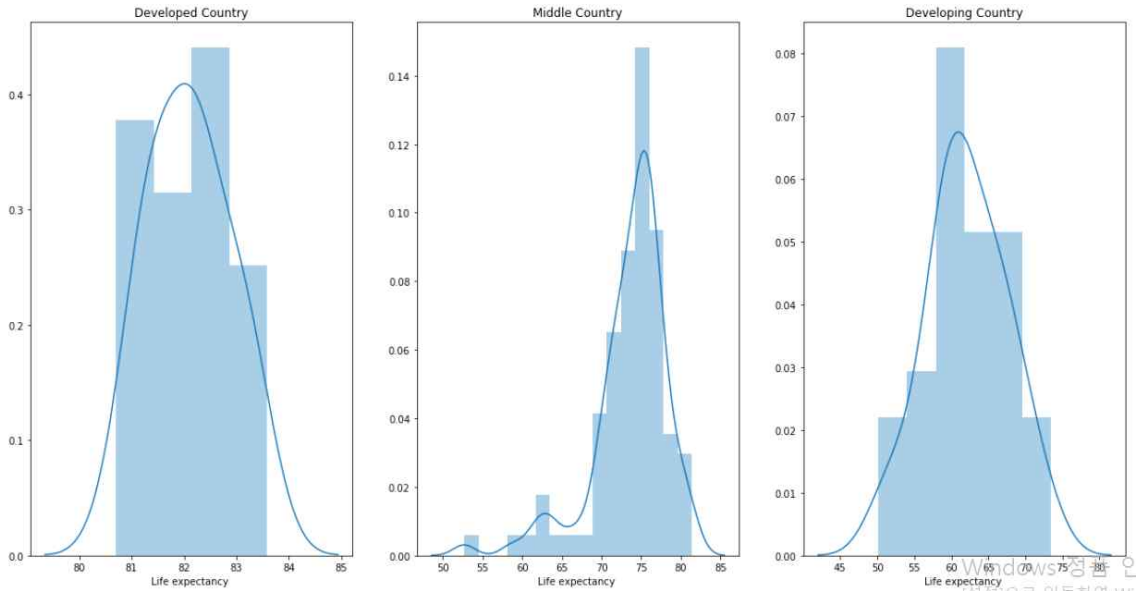
이러한 방법으로 모델을 적합시키면 위의 **그림 <2>** 와 같은 형태의 변수들을 각 수준에 따라 잘 적합할 수 있을 것이라 생각하였다. 물론 저와 같은 형태의 변수를 적합시킬 때 사용할 수 있는 변수변환의 방법도 사용하여 더 좋은 결과를 보여주는 방식을 택할 것이다.

그리고 이렇게 선진국과 후진국으로 분류한 후에 나머지 국가들을 중진국으로 분류하였다. 한 가지 예외 국가로써, 중진국으로 분류되었던 적도기니를 제외했다. 적도기니의 1인당 GDP는 15000달러로 상당히 높지만 기대수명은 60세가 채 안될 정도로 보건이나 기타 생활여건은 낮다. 이는 유전에 의해 국민들의 생활수준에 비해 1인당 GDP가 지나치게 높은 경우로써 일반적이지 않았다.

나. 사용데이터 국가들(Rows)

결과적으로 위에서 제외한 국가 적도기니와, health indicator 데이터에 결측치가 존재하여 제외한 10개국가(cabo verde) 을 제외하고 중간과제에서보다 11개적은 141개 국가의 데이터를 사용하게 되었다. (선진국 21개국가, 중진국 86개국 ,후진국 34개국)

또한 위의 기준에 따라 국가들을 분류한 결과 선진국 21개국가, 중진국 86개국가, 후진국은 34개의 국가들로 분류되었다. 국가수준에 따라 국가별 기대수명 그래프를 보았을 때도 확연하게 기대수명의 분포에서 차이가 나는 것을 확인할 수 있었다. 이를 잘 구분하여 모델링한다면 좋은 결과를 얻을 수 있을 것이라고 예상했다.



<그림 3> 선진국, 중진국, 후진국 별 기대수명의 분포
(Seaborn package의 distplot 사용)

제 2장 본론

제 1절 변수 선택

가. 변수선택 기준

술, 담배, 비만등 (생활습관)	질병	국가의 부	교육	기타
alcohol per capita	Cancers (%)	GDP_per capita	Schooling	Total expenditure
Share of deaths from smoking (%)	Diarrhea & common infectious diseases (%)			Adult Mortality
obesity	Nutritional deficiencies (%)			
	pneumonia-death-rates			
	Diabetes, blood, & endocrine diseases (%)			
	Liver disease (%)			
	Diphtheria			
	HIV/AIDS_a			
	thinness 1-19 years			
	Cardiovascular diseases (%)			

그림 <4> 현재 모델에 적합시킬 수 있는 변수후보들

선진국 , 중진국 ,후진국으로 나누어 모델을 적합시키기 위해 선택해야할 변수들을 추려내야했다. 변수를 고르기에 앞서 변수 선택에 몇 가지 기준을 정하였다.
첫 번째, 모델에 1인당 GDP와 'Schooling' 은 반드시 포함하기로 하였다. 그 이유

는 가장 먼저 두 변수가 회귀모델을 만들 때 좋은 성능을 보여줬기도 했지만 두 변수가 대표성을 가지고 있다고 생각하였기 때문이다. 예를 들어, 1인당 GDP는 선진국, 중진국, 후진국등의 구별에 쓰이는 중요한 변수로 쓰이기도 하면서 국가의 부의 정도를 나타내는 좋은 측도였다. 그리고 'Schooling' 은 평균 교육년수로써 교육의 정도와 기대수명의 관계를 나타내는 좋은 변수이다

현재 내가 가지고 있는 데이터들의 변수들은 대부분은 질병과 관련된 변수이다. 담배, 술, 비만등과 관련된 변수는 생활습관이라는 대표성을 가진다고 묶어 보았다. 모델의 성능이 나쁘지 않다면 질병을 대표하는 변수, 생활습관, 국가의 부 등등 최대한 다양한 특성을 가진 변수들을 이용하는 것이 좋을 것이라고 생각했다. 그래서 한 특성에 속한 변수들이 좋은 성능을 보여주더라도 각 특성에서 최대 2개까지 사용하기로 하였다.

나. 산점도를 통한 변수선택

우선 가지고 있는 변수의 개수가 너무 많았기 때문에 기대수명과 변수와의 산점도를 그려본 후 선형성이 보이지 않거나 어떠한 패턴도 보이지 않는 변수들을 제거하였다. 특히 , 질병과 관련한 변수들은 그 수가 많기 때문에 패턴이 뚜렷하지 않거나 선형성이 뚜렷하지 않은 데이터등 까다로운 기준으로 제거하였다. 산점도를 그려본 결과 다음과 같은 변수들이 남게 되었다. (산점도의 개수가 너무 많으므로 그림은 **부록<2>**를 참고) 변수선택의 결과 중요했던 점을 위주로 간추려 설명하겠다.

술, 담배, 비만등 (생활습관)	질병	국가의 부	교육	기타
alcohol per capita	Cancers (%)	GDP_per capita	Schooling	Total expenditure
Share of deaths from smoking (%)	Diarrhea & common infectious diseases (%)			Adult Mortality
obesity	Nutritional deficiencies (%)			
	pneumonia-death-rates			
	Diabetes, blood, & endocrine diseases (%)			
	Liver disease (%)			
	Diphtheria			
	HIV/AIDS_a			
	thinness 1-19 years			
	Cardiovascular diseases (%)			

그림 <5> 산점도를 통하여 포함시키지 않은 변수들 (빨간색으로 표시)

나-1.Adult Mortality 변수를 제거 하기로 하였다.

기대수명은 각 나이의 사망률을 이용하여 구해진다. 그런데 adult morality는 15~60세의 사망률을 표시한 지표로써 원래 목표하고자 했던 것 중 하나인 '직접적인 사망률을 모르더라도 기대수명을 구할 수는 없을까?' 에 부합하지 않는다고 생각하였다. 비록 중간과제까지는 모델에 포함되었던 변수이고 선형성이 뚜렷한 변수이지만 제외하기로 하였다.

나-2. Total expenditure와 질병의 빨간색 변수들 5개를 제거했다.

가장 오른쪽의 Total expenditure는 중간과제에서도 채택되지 않은 변수로써 기대수명과의 선형성이 강하지 않았으며 질병에 있는 5개의 빨간색 변수들은 나머지 변수들보다 기대수명과의 선형성이 좋지 않았으므로 굳이 사용할 이유가 없었다. 지우지 않은 변수들로만 회귀모형을 구축해도 충분히 좋은 결과가 나올 것 이라고 생각했다.

나-3. alcohol데이터는 남겨두되 obesity 데이터는 삭제하였다.

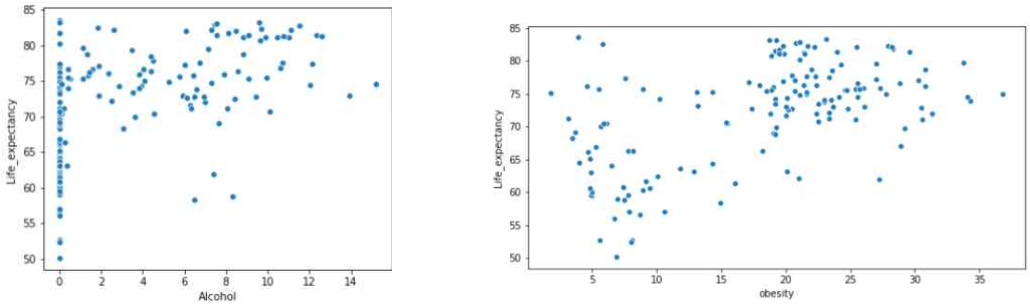


그림 <6> Alcohol과 기대수명 ,Obesity와 기대수명의 산점도

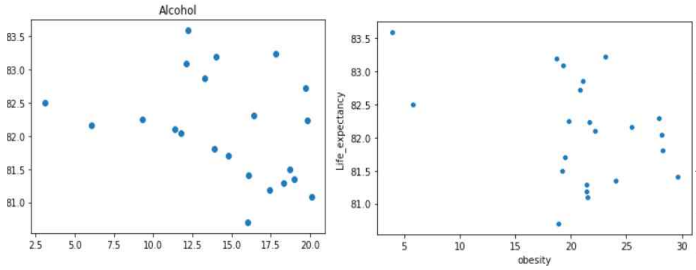


그림 <7> 선진국들의 Alcohol과 기대수명 ,Obesity와 기대수명의 산점도

그림 <6>을 봤을 때 이 두 변수들은 기대수명과의 선형성이 그렇게 좋다고 말할 수는 없다. Alcohol의 경우 0주변에 몰려있고 Obesity의 경우 obesity=15를 주변으로 데이터들이 군집되 있는 것 같은 형태를 보여준다.

하지만 이 2변수를 선진국 사이들에서 그려보면 (그림<7>) 알코올은 선진국들 사이에서 기대수명에 어느 정도 연관이 있는 것 처럼 보였지만 obesity는 그렇지 않았다. 따라서 선형성이 아주 좋았던 것은 아니지만 Alcohol을 사용한다면 선진국수준의 예측변수들을 모델링할 때 유용할 것이라고 생각하여 남겨두기로 하였다.

추가적으로 선택된 변수들에 대하여 각 수준별(선진국,중진국,후진국) 기대수명과의 산점도를 그려보았다(**부록 <3>**). 이 산점도들은 후에 각 수준별로 데이터들의 추세가 많이 다를 경우 이 산점도들을 고려하여 모델링 하기로 하였다. 예를 들면, Cancers 의 경우 각 수준별 데이터들의 추세가 다르다. 이를 다음과 같이 각각의 추세로 모델링하는 방안(**그림 <2>**)과 위에서 서술했듯이 변수에 변환을 하여 더미변수를 쓰지 않고 적합하는 방식, 가령 **그림 <8>** 에서는 비슷한 형태의 함수인 $y=\sqrt{x}$ 형태로 적합시키는 방식과 비교하여 더 좋은 모델을 사용하고자 한다.

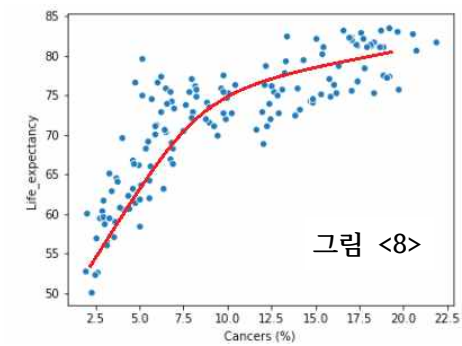


그림 <8>

다. 상관관계 및 다중공선성 확인

	Life_expectancy	GDP_per	Schooling	Cancers	Diarrhea_common	Nutrition	pneumonia	Alcohol	smoking
Life_expectancy	1.000000	0.613092	0.819755	0.834467	-0.870830	-0.813301	-0.849114	0.348165	0.717595
GDP_per	0.613092	1.000000	0.610879	0.608508	-0.463297	-0.497715	-0.428163	0.351408	0.391707
Schooling	0.819755	0.610879	1.000000	0.813128	-0.809537	-0.786568	-0.727467	0.547311	0.683207
Cancers	0.834467	0.608508	0.813128	1.000000	-0.772298	-0.776673	-0.716999	0.623759	0.783977
Diarrhea_common	-0.870830	-0.463297	-0.809537	-0.772298	1.000000	0.882993	0.893963	-0.394486	-0.751449
Nutrition	-0.813301	-0.497715	-0.786568	-0.776673	0.882993	1.000000	0.795233	-0.437109	-0.741014
pneumonia	-0.849114	-0.428163	-0.727467	-0.716999	0.893963	0.795233	1.000000	-0.334507	-0.709752
Alcohol	0.348165	0.351408	0.547311	0.623759	-0.394486	-0.437109	-0.334507	1.000000	0.420646
smoking	0.717595	0.391707	0.683207	0.783977	-0.751449	-0.741014	-0.709752	0.420646	1.000000

그림 <10> 변수들의 상관행렬

다음으로 설명변수와 예측변수의 상관성과 선명변수들의 상관성을 보기위해 상관행렬을 만들어보았다. 파란색네모를 봤을때 Alcohol을 제외하고 기대수명과 관계가 있어 보이는 변수들을 골랐던 만큼 상관관계가 높게 측정되었다. 변수들 사이의 상관관계를 살펴보면 질병과 관련된 변수들 사이에 상관관계가 높게 나와 연구초반에 걱정했던 것처럼 다중공선성이 발생할 확률이 높아보였으며 많이 사용하지 않는 것이 좋다고 생각하였다. VIF를 구해본결과 Diarrhea_common변수가 9.454744로 질병변수들을 대체할 다른 변수들도 충분히 있으므로 제거하는편이 낫다고 판단하였다.

```
vif(lm_vif)
     GDP_per     Schooling       Cancers DIarrhea_common     Nutrition     pneumonia
    1.834020      4.538106      5.839124      9.454744      5.241927      5.129227
     Alcohol       smoking
    1.822084      3.221882
```

그림 <11> 변수들의 VIF 값

제 2절 Model Building

가. GDP변수와 Schooling변수

모델을 설정하기에 앞서 설명한대로 1인당 GDP와 Schooling변수는 반드시 포함하기로 하였으므로 전진 선택법을 통하여 GDP와 Schooling 변수는 포함하여 모델을 선택하기로 하였다. 모델선택기준으로써는 수정 결정계수와 AIC 값을 참고하기로 하였다. 가장 처음 GDP와 Schooling을 포함한 모델을 만들어보았다.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.520e+01  2.023e+00  22.345  < 2e-16 ***
GDP_per     6.965e-05  2.319e-05   3.004  0.00317 **
Schooling   1.953e+00  1.640e-01  11.906  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.393 on 138 degrees of freedom
Multiple R-squared:  0.6921,    Adjusted R-squared:  0.6877
F-statistic: 155.1 on 2 and 138 DF,  p-value: < 2.2e-16
```

그림 <12> 기대수명~ GDP +Schooling 모델의 결과 와 AIC 값

```
> AIC(lm1)
[1] 822.4805
```

여기서 국가수준에 따라 분리하여 가능한 모든 조합에서 모델링을 해보기로 하였다. 예를 들자면 다음과 같은 방식으로 각 수준에 따라 기울기가 달라지도록 모델링하였다.

(ex) 기대수명= GDP+ factor(선진국일 때) * GDP+ factor(중진국일때)*GDP + Schooling **GDP에서 모든수준으로 나눈 모델**

기대수명 = 1인당GDP+ factor(선진국) * 1인당GDP +Schooling
GDP에서 선진국으로만 분류한모델

.....

기대수명 = 1인당GDP+ factor(선진국)*1인당 GDP+factor(중진국)*1인당GDP + Schooling + factor(선진국)*Schooling + factor(중진국)*Schooling
GDP,Schooling 모두에서 수준을 고려한 모델

하지만 결과는 다음과 같았다.

```
> lm2=lm(Life_expectancy~GDP_per+GDP_per*factor(Developed_country)+GDP_per*factor(Middle_country)+Schooling,data=raw4)
> summary(lm2)
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              48.483764   2.351215  20.621  < 2e-16 ***
GDP_per                                   0.002669   0.001539   1.734   0.0852 .
factor(Developed_country)1               14.525856   3.484263   4.169 5.46e-05 ***
factor(Middle_country)1                   8.581860   2.100209   4.086 7.51e-05 ***
Schooling                                 1.114798   0.206709   5.393 3.04e-07 ***
GDP_per:factor(Developed_country)1 -0.002665   0.001540  -1.731   0.0858 .
GDP_per:factor(Middle_country)1    -0.002544   0.001536  -1.657   0.1000 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.975 on 134 degrees of freedom
Multiple R-squared:  0.7553,    Adjusted R-squared:  0.7443
F-statistic: 68.93 on 6 and 134 DF,  p-value: < 2.2e-16
```

그림 <13> 기대수명~ GDP+ GDP(선진국일 때)+GDP(중진국일 때)+ Schooling

수준별로 적합시킨 모델의 예로 그림<13>을 가져왔다. 결과를 보면 모델의 수정계수 값이 올라가더라도 **각 변수들의 T-value의 값들이 크게 감소하여** 유의수준 0.05를 넘는 경우가 발생하였다. VIF값을 확인해보니 값이 10이상을 넘어 심각한 수준으로 올라가는 것을 확인하였다 (그림<14>).

```
> vif(lm2)
                     GDP_per          factor(Developed_country)          factor(Middle_country)                       Schooling
                 8587.698217                         13.733482                          9.365789                        3.094338
GDP_per:factor(Developed_country)    GDP_per:factor(Middle_country)
                 9562.042791                       1346.462620
```

그림 <14> 그림 <13>의 VIF 값

이번엔 GDP 대신 Schooling을 국가수준에 따라 모델에 적합시켜보았다.

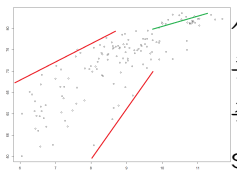
```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.295e+01  2.038e+00  21.079  < 2e-16 ***
GDP_per                          4.989e-05  3.447e-05   1.447 0.150125
Schooling                        2.143e+00  1.650e-01  12.988  < 2e-16 ***
factor(Developed_country)1       3.783e+01  1.009e+01   3.751 0.000260 ***
Schooling:factor(Developed_country)1 -2.228e+00  5.836e-01  -3.818 0.000204 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.206 on 136 degrees of freedom
Multiple R-squared:  0.7219,    Adjusted R-squared:  0.7137
F-statistic: 88.27 on 4 and 136 DF,  p-value: < 2.2e-16

> vif(lm7)
                           GDP_per                          Schooling          factor(Developed_country)
                          3.846608                           1.761237                          102.796059
Schooling:factor(Developed_country)
                          99.785261
```

그림 <15> 기대수명~ GDP+ Schooling + Schooling(선진국일 때)*GDP

그 결과 수정결정계수값이 상승하였지만 변수들의 t-value값이 유의하지 않았으며 VIF값을 확인해보니 역시 다중공선성이 발견됨을 확인할 수 있었다.

그 어떤 조합으로 모델을 구성해도 국가수준별 모델링 방식은 다중공선성을 일으켰고 결론적으로, 국가수준별 모델링 방식은 사용할 수 없다 고 결론을 내렸다.

그러므로 $Y=\sqrt{x}$ 의 형태를 가졌던 1인당 GDP 변수를 $\sqrt{1\text{인당 }GDP}$ 형태로 변수변환하여 모델에 넣어보았고 다음과 같이 수정결정계수값이 상승하고 AIC값도 하락했으며 변수들의 유의성도 만족하였다. 비슷한 형태의 다른 변환 (log변환, $x^{1/3}$등)을



사용하게 되면 왼쪽 그림과 같이 좌측 아랫방향으로 벌어지는 형태를 보여주게되며 수정결정계수의 증가량에서 $\sqrt{1\text{인당 }GDP}$ 변환보다 좋은 모습을 보여주지 못하였다. 따라서 기대수명= $\sqrt{1\text{인당 }GDP}$ + schooling 모형을 이 단계에서는 택하기로 하였다.

```
> lm9=lm(Life_expectancy~sqrt(GDP_per)+Schooling,data=raw4)
> summary(lm9)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    46.858145   2.039835  22.972  < 2e-16 ***
sqrt(GDP_per)   0.033270   0.007937   4.192 4.92e-05 ***
Schooling       1.654863   0.190738   8.676 1.02e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.271 on 138 degrees of freedom
Multiple R-squared:  0.709,     Adjusted R-squared:  0.7048
F-statistic: 168.1 on 2 and 138 DF,  p-value: < 2.2e-16
> AIC(lm9)
[1] 814.5125
```

그림 <16> 기대수명~ sqrt(GDP) +Schooling 모델 결과와 AIC값

나. 질병변수

질병변수 pneumonia , nutrition deficiency , cancers를 이용하여 모델을 적합해봤다. 우선 세 변수들 중 pneumonia를 더했을 때 모델의 수정 결정계수값과 AIC 값이 크게 향상 되었다.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    64.926437   2.305939  28.156  < 2e-16 ***
sqrt(GDP_per)   0.030207   0.005948   5.078 1.22e-06 ***
Schooling       0.661121   0.171509   3.855 0.000177 ***
pneumonia      -0.104097   0.009956 -10.456  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.197 on 137 degrees of freedom
Multiple R-squared:  0.8382,    Adjusted R-squared:  0.8346
F-statistic: 236.5 on 3 and 137 DF,  p-value: < 2.2e-16

> AIC(lm11)
[1] 733.7933
```

그림 <17> 기대수명~ sqrt(GDP) + Schooling + pneumonia

```
> lm12=lm(Life_expectancy~sqrt(GDP_per)+Schooling+pneumonia+Nutrition,data=raw4)
> summary(lm12)

Call:
lm(formula = Life_expectancy ~ sqrt(GDP_per) + Schooling + pneumonia +
    Nutrition, data = raw4)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2622  -1.9745   0.1001   2.1578   8.8027

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.636836   2.609396  25.920  < 2e-16 ***
sqrt(GDP_per)  0.028675   0.005917   4.846 3.39e-06 ***
Schooling      0.519178   0.182029   2.852  0.00502 **
pneumonia     -0.091201   0.011551  -7.896 8.62e-13 ***
Nutrition     -0.663909   0.312234  -2.126  0.03528 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.156 on 136 degrees of freedom
Multiple R-squared:  0.8434,    Adjusted R-squared:  0.8388
F-statistic: 183.1 on 4 and 136 DF,  p-value: < 2.2e-16
```

```
> lm13=lm(Life_expectancy~sqrt(GDP_per)+Schooling+pneumonia+Cancers,data=raw4)
> summary(lm13)

Call:
lm(formula = Life_expectancy ~ sqrt(GDP_per) + Schooling + pneumonia +
    Cancers, data = raw4)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8360 -1.7632  0.2398  1.9173  9.5329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.479380   2.182332  30.004  < 2e-16 ***
sqrt(GDP_per)  0.022826   0.005889   3.876 0.000165 ***
Schooling      0.349048   0.178348   1.957 0.052382 .
pneumonia     -0.091068   0.009907  -9.193 5.84e-16 ***
Cancers        0.361498   0.086362   4.186 5.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.02 on 136 degrees of freedom
Multiple R-squared:  0.8566,    Adjusted R-squared:  0.8524
F-statistic: 203.2 on 4 and 136 DF,  p-value: < 2.2e-16
```

그림 <18>= 그림 <17> 모델에 각각 nutrition 과 cancers를 더한 모델

여기서 nutrition과 cancers를 각각 더한 모델을 구해보았다. 변수들을 더하여 모델을 적합해본결과 모델이 향상되긴 하였지만 큰 변화가 있는 것은 아니었다. nutrition을 더한 모델의 경우 수정결정계수의 값은 고작 0.0043 증가하였다. 게다가 오른쪽 cancers를 더한 모델의 경우 schooling의 t-value에 대한 유의확률 값이 0.05를 넘게 되었다. 그래서 cancer를 더한 모델은 사용할 수 없으며 nutrition을 더한 모델은 우선 보류하고 이 단계에서는 pneumonia만을 더한 모델을 남기고 모델을 최종적합한 후 nutrition만을 한번 더 고려하기로 하였다.

다. 생활습관 변수

마지막으로 alcohol변수와 share of deaths from smoking 변수를 추가해보았다. 그림<19>. 그 결과 Alcohol변수를 추가하였을 때 수정결정계수값은 0.8406 , AIC값이 729.56까지 감소하여 smoking을 추가 하였을때보다 더 좋은 결과를 보여주었다. **(그림 <19> 왼쪽그림)** 따라서 Alcohol을 추가하기로 결정하였고 이 모델에서 아까 고려해보기로 한 Nutrition변수를 더해보았다. **(그림 <19> 오른쪽그림)**

```
> lm14=lm(Life_expectancy~sqrt(GDP_per)+Schooling+pneumonia+Alcohol,data=raw4)
> summary(lm14)

Call:
lm(formula = Life_expectancy ~ sqrt(GDP_per) + Schooling + pneumonia +
    Alcohol, data = raw4)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3419 -1.9264  0.1361  1.8962  9.6707

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.772168   2.311209  27.593  < 2e-16 ***
sqrt(GDP_per)  0.030529   0.005841   5.227 6.35e-07 ***
Schooling      0.834163   0.182272   4.576 1.05e-05 ***
pneumonia     -0.101372   0.009836 -10.306  < 2e-16 ***
Alcohol       -0.124699   0.050303  -2.479   0.0144 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.138 on 136 degrees of freedom
Multiple R-squared:  0.8452,    Adjusted R-squared:  0.8406
F-statistic: 185.6 on 4 and 136 DF,  p-value: < 2.2e-16

> AIC(lm14)
[1] 729.5618
```

```
> lm15=lm(Life_expectancy~sqrt(GDP_per)+Schooling+pneumonia+Alcohol+Nutrition,data=raw4)
> summary(lm15)

Call:
lm(formula = Life_expectancy ~ sqrt(GDP_per) + Schooling + pneumonia +
    Alcohol + Nutrition, data = raw4)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8337 -2.0231  0.1138  2.0676  7.8028

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.670521   2.575434  25.887  < 2e-16 ***
sqrt(GDP_per)  0.028865   0.005784   4.990 1.83e-06 ***
Schooling      0.691171   0.188914   3.659 0.000363 ***
pneumonia     -0.086939   0.011400  -7.626 3.87e-12 ***
Alcohol       -0.134438   0.049620  -2.709 0.007617 **
Nutrition     -0.732028   0.306235  -2.390 0.018210 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.085 on 135 degrees of freedom
Multiple R-squared:  0.8515,    Adjusted R-squared:  0.846
F-statistic: 154.8 on 5 and 135 DF,  p-value: < 2.2e-16

> AIC(lm15)
[1] 725.7166
```

그림 <19> Alcohol만을 추가한 모델 / Alcohol, Nutrition을 모두 추가한 모델

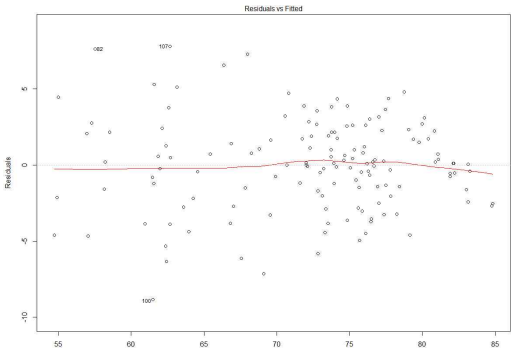
그 결과 수정결정계수값은 0.846 으로 0.0054만큼 증가하고 AIC값도 4만큼 감소하였다. 이는 유의미한 성능의 향상이라 생각하였고 각 변수들의 t_value값을 확인한 결과 변수들도 모두 유의하였다. 또한 VIF값을 통하여 다중공선성을 확인해보아도 큰 문제가 되지 않았다 (10 이하).

```
> vif(lm15)
sqrt(GDP_per)     Schooling     pneumonia       Alcohol     Nutrition
     2.322450      4.289706      2.996774      1.455937      3.707355
```

그림 <20> Alcohol nutrition을 추가한모델의 vif 값

3절 회귀모델 가정 성립 확인

1. 등분산성



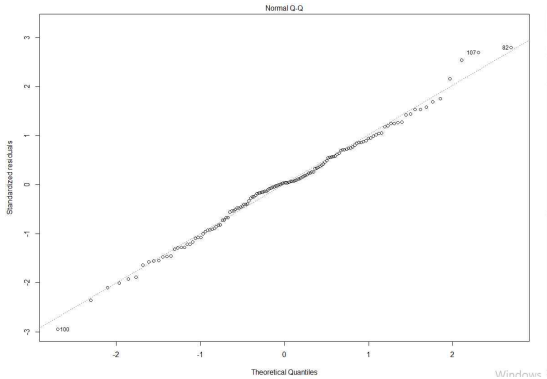


그림 <21> Residuals vs fitted_values　　Normal Q-Q plot

Residuals vs fitted_values plot을 살펴본 결과 아닌 점들이 발견되긴 하였지만 대체로 잔차들이 -5에서 5사이에 분포하였다. 따라서 등분산성을 만족한다고 결론을 내렸다.

2. 정규성

```
> shapiro.test(lm19$residuals)

        Shapiro-Wilk normality test

data:  lm19$residuals
W = 0.99425, p-value = 0.8482
```

정규성을 확인하기 위해 Normal Q-Q plot을 그려본 결과 정규성을 잘 만족하는 것처럼 보였다 (그림 <21> 오른쪽 그래프). 추가적으로 샤피로 검정을 진행한 결과 유의수준 0.05에서 귀무가설을 기각할 수 없으므로 정규분포를 만족한다고 결론을 내렸다.

3 독립성

```
> durbinWatsonTest(lm19)
 lag Autocorrelation D-W Statistic p-value
   1    -0.007791247       2.004056   0.972
 Alternative hypothesis: rho != 0
```

마지막으로 독립변수들 간에 독립성을 확인하기 위하여 더빈-왓슨 검정을 하였다. 검정결과 검정통계량이 2에 상당히 가깝게 나왔고 p-value도 0.972로써 상당히 높게나와 독립성을 만족한다고 결론을 내렸다.

결론적으로, 모델은 등분산성, 정규성, 독립성을 모두 만족하였다. 최종적으로 이 모델을 선택하기로 하였다.

기대수명= 66.670521 + 0.028865*sqrt(1인당_GDP) + 0.691171*schooling(평균 교육연수) + -0.086939* pneumia(폐렴으로의한 DALY) + -0.732028 * Nutrition (인구중 영향결핍인구비율) + -0.134438*(순수 알코올 섭취량, {단위=Liter})

제 3절 변수의 중요도 및 모델 평가.

가. 변수의 중요도

```
> summary(lm17)

Call:
lm(formula = Life_expectancy ~ GDP_per + Schooling + pneumonia
    Alcohol + Nutrition, data = raw5)

Residuals:
     Min       1Q   Median       3Q      Max
-1.12377 -0.25737  0.01448  0.26303  0.99263

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.019e-17  3.305e-02   0.000 1.000000
GDP_per      2.523e-01  5.055e-02   4.990 1.83e-06 ***
Schooling    2.514e-01  6.870e-02   3.659 0.000363 ***
pneumonia   -4.379e-01  5.742e-02  -7.626 3.87e-12 ***
Alcohol     -1.084e-01  4.002e-02  -2.709 0.007617 **
Nutrition   -1.527e-01  6.387e-02  -2.390 0.018210 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3925 on 135 degrees of freedom
Multiple R-squared:  0.8515,    Adjusted R-squared:  0.846
F-statistic: 154.8 on 5 and 135 DF,  p-value: < 2.2e-16
```

그림 <22> sclae함수를 적용한 후 모형을 만든 모델의 결과

기존 회귀모델에서는 변수들의 단위가 통일되어있지 않기 때문에 변수끼리 비교를 하기 위해서는 변수들을 통일 시키는 과정이 필요하였다. 그래서 각 변수들을 R의 scale함수를 통하여 정규화시키고 회귀모델을 돌려보았다. 회귀계수들의 유의성이나 수정결정계수들의 값은 당연하게도 통일화하기전 모델과 같은 값이 나오게 되었다.
단순 절대값만 살펴보았을 때 Pneumonia가 한 단위가 증가할 때 약 4.379 만큼 기대수명에 영향을 주는 변수로써 가장 영향력이 큰 변수로 나타났다.

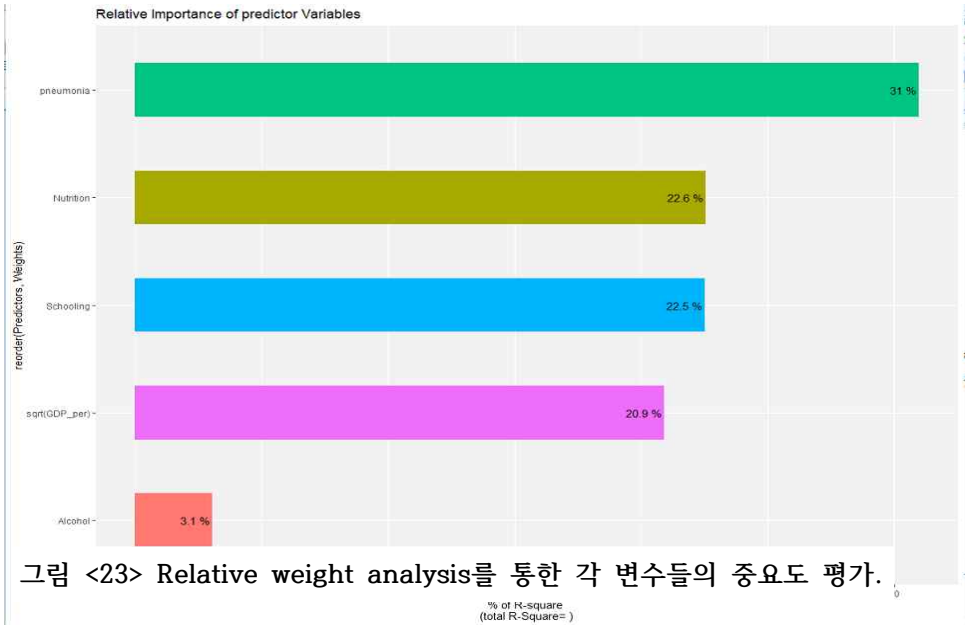
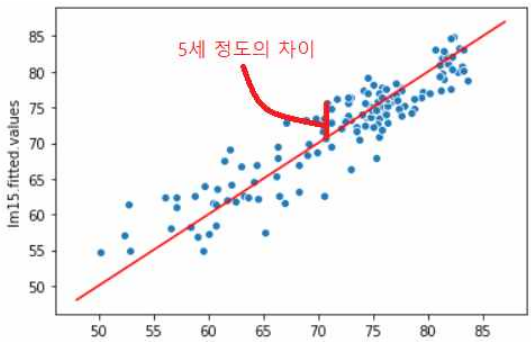


그림 <23> Relative weight analysis를 통한 각 변수들의 중요도 평가.

또한 Relative weight analysis를 통하여 각 변수들이 R-square값을 얼만큼 설명해주고있는지를 알아보았다 (그림 <23>). Relative weight analysis는 쉽게 설명하자면 가능한 모든 하위모델에서 예측변수를 추가하였을 때 얻은 R제곱의 값의 증가량의 평균을 구한 것이다. 그리고 각 변수들의 전체합을 100%를 기준으로 **(자세한설명과 사용코드는 각주와 부록참조)**[2] 마찬가지로 pneumonia가 약 31프로로써 가장 높았으며 Alcohol변수가 3.1프로로써 가장 낮았다. 상관행렬을 만들어봤을 때 가장 상관계수가 높았던 pneumonia가 모델에 가장 많이 기여를 했고 다소 상관계수 값이 떨어졌던 Alcohol의 경우 모델의 기여도가 낮게 나온 것이다. 이를 통하여 예측변수와 상관성이 높은 변수를 넣어야 좋은 모델이 나올 확률이 높다는 것을 확인할 수 있었다.

나. 모델의 성능평가



이 모델은 등분산성을 만족하였지만 residual vs fitted 그래프에서 확인할 수 있었듯이 잔차들이 5에서 -5정도 범위 안에서 무작위로 찍혀있는 형태였기 때문에 분산이 크다고 할 수 있었다. 또한 기대수명의 Rmse($\sqrt{\text{차의 제곱들의 평균}}$) 값을 구해보았다. Rmse 값은 3.01886 값이 나왔다.

그림 <24> 실제값 vs 예측값의 산점도

이제 이 모델을 실제 기대수명을 예측하는데 사용해보았다.

	Alcohol	Nutrition	pneumonia	Schooling	GDP_per	기대수명
Andorra	12.4	0.340551	18.23608013	11.4	42,300.30	82.65
Czechia	11.8	0.298047	16.11077301	14	19744.6	78.82
Djibouti	10.4	3.570635	80.95906449	9.2	2466.02	63.17
Morocco	0.5	0.509331	7.627206166	10	3171.7	75.48
Bahrain	1.8	1.38523	19.13544645	14.5	24989.38	76.624
Bhutan	0.7	4.422314	31.28634726	12.5	2652.215	70.046
Papua Ne	2.3	2.976249	114.6050038	10	2920.784	63.181
한국	12.3	0.818717	18.07324735	12	27811	81.72
미국	8.82	0.90013	24.22829793	12.9	55033	78.84

그림 <25> 실제 데이터 값

	fit	lwr	upr
1	71.13472	64.80243	77.46701
2	73.18254	66.87675	79.48833
3	62.02881	55.83413	68.22348
4	72.56071	66.15942	78.96200
5	73.83060	67.47836	80.18283
6	69.32333	62.90665	75.74000
7	61.20701	54.93686	67.47717
8	71.21110	64.93305	77.48915
9	71.72686	65.47541	77.97831

모델의 예측 기대수명과 95%신뢰구간값

2) https://medium.com/analytics-vidhya/johnsons-relative-weights-analysis-implementation-with-javascript-d85393c0bbb4#:~:text=The%20Johnson's%20Relative%20Weights%20(JRW,are%20correlated%20to%20each%20other.

데이터를 가져올 때 모든 변수후보들(현재 4개의 변수 외에 변수 후보에 올랐던 obesity, cacner 등등 모두포함) 중에 완벽히 값들이 다 존재하였던 141개의 국가를 사용하였다. 하지만 최종모델로 적합된 모델의 4개의 설명변수에서 완벽히 값들이 다 존재하는 국가들은 148개 국가로 다행히 148-141= 7 개의 데이터들을 구할 수 있었다. 여기에 상대적으로 검색을 통하여 쉽게 구할 수 있었던 한국과 미국의 자료값을 이 모델에 넣어 보았다. 테스트 데이터가 9개밖에 안되었지만 이 결과 만을 가지고 판단해 볼 때 기대수명이 낮은 국가들(djibouti , bhutan , papua new guinea) 의 경우 예측값이 실제 기대수명과 가까웠지만 Andorra,한국과 같이 기대수명이 높은 국가들의 경우 오차가 10세 가까이 날 만큼 오차가 크게 나타났다. 또한 9개 국가들 중 6개 국가들은 실제 기대수명이 95프로 신뢰구간 안에 포함되었지만 Andorra, 한국 , 미국 세 나라는 실제 기대수명이 95프로 신뢰구간에도 포함되지 못하였다. 물론 테스트데이터가 9개밖에 안되지만 대체적으로 기대수명이 낮은 국가에서 정확도가 높아지고 높은 국가에서 정확성이 떨어지는 이유에 대해 생각해보았다.

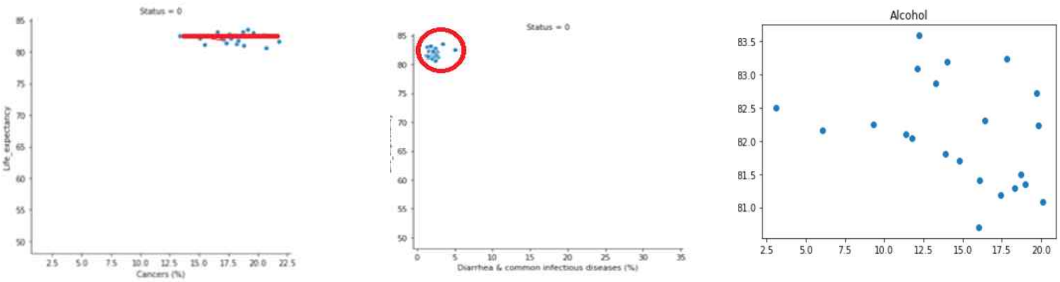


그림 <26> 선진국에서의 설명변수와 기대수명과의 관계

역시 가장 큰 이유는 기대수명이 높은 국가들에서 뚜렷한 패턴을 보여주는 변수를 찾지 못한 것이 가장 큰 원인이라고 생각하였다. 아무리 선형성이 뚜렷한 변수라 하더라도 기대수명이 높은 국가들에서는 그림<26>의 왼쪽, 가운데 그림을 보면 확인할 수 있듯이 기울기가 0에 가깝거나 패턴이 없이 그냥 군집되있는 형태를 보여주었다. 이를 해결하고자 그나마 가능성이 있어보였던 Alcohol을 고려해보았었지만 효과가 적었던 것으로 추측 된다 (그림 <26> 오른쪽 그림).

3장 결론

결론 및 해석

최종모델:

기대수명= 66.670521 + 0.028865*sqrt(1인당_GDP) + 0.691171*schooling(평균 교육연수) + -0.086939* pneumia(폐렴으로의한 DALY) + -0.732028 * Nutrition (인구중 영향결핍인구비율) + -0.134438*(순수 알코올 섭취량, {단위=Liter})

지금까지 141개의 국가를 통하여 1인당 GDP, 평균교육연수, pneumonia의 DALYS ,1인당 연간 알콜섭취량을 통하여 기대수명의 값을 표현하고 예측할 수 있는 다중선형회귀모형을 만들어보았다. 원래 목표하고자 했던 국가수준별 분류방법은 다중공선성을 발생시켜 사용할 수 없던 점은 아쉬웠지만 그 방법을 사용하지 않고도 회귀모델의 가정을 잘 만족시키며 기대수명을 대표성을 가지는 변수들로 예측할 수 있는 모델을 만들 수 있었다.

기대수명이 높은 국가들은 잘 예측하지 못하고 모델의 RMSE 값이 약 3으로 정확도가 매우 높다고 할 수 는 없지만, 한 나라의 기대수명의 대략적인 경향을 예측하기에는 충분히 괜찮은 모델이라고 생각했다. 회귀모델에서는 잔차의 등분산성을 가정할 뿐 그 분산이 몇 이하여야한다는 조건은 없다. 회귀모델의 장점은 예측력이 아니라 설명변수와 종속변수의 경향을 잘 보여주며 선형이기 때문에 해석이 용이한 점이라고 생각한다. 이 모델은 5가지 변수만을 이용하여 간단한 선형함수로 쉽고 대략적으로 기대수명값을 예측해 볼 수 있는게 큰 장점이라고 생각한다.

모델을 변수들의 대표성을 고려하여 만들었지만 가장 중요한 변수는 Pneumonia 이었다. 다른 변수들에 비하여 한 단위 증가할 때 기대수명에 가장 큰 영향을 주는 변수였으며 결정계수값의 대부분을 설명해주는 변수였다. pneumonia는 EDA에서부터 기대수명과의 선형성이 매우 좋았던 변수로써 좋은 회귀모형을 적합시키기 위하여 종속변수와 상관성이 높은 설명변수를 사용해야한다는 것을 알 수 있었다.

pneumonia는 질병 또는 보건성 을 대표하는 지표로써 고른 것이다. 실제로 pneumonia는 폐렴으로 선진국등에서는 거의 해결된 질병이지만 후진국에서 크게 문제가 되는 질병으로써 후진국병으로 불릴만큼 그 나라의 보건성을 잘 보여주는 질병중에 하나이기도 하다. 다른 지표들 (국가의부, 교육의정도, 생활습관) 이 비록 서로 연관성이 있는 것은 사실이다. 예를 들어, 국가의 부를 올린다면 돈이 증가하기 때문에 의료, 교육등 다양한 변수들에서 값이 상승할 것 이다. 하지만 질병이 기대수명에 가장 큰 영향을 미치는 척도로써 질병에 대한 직접적인 개선이 다른 변수들에 비해 기대수명을 늘리는데 가장 큰 영향을 줄 것으로 확인할 수 있었다.

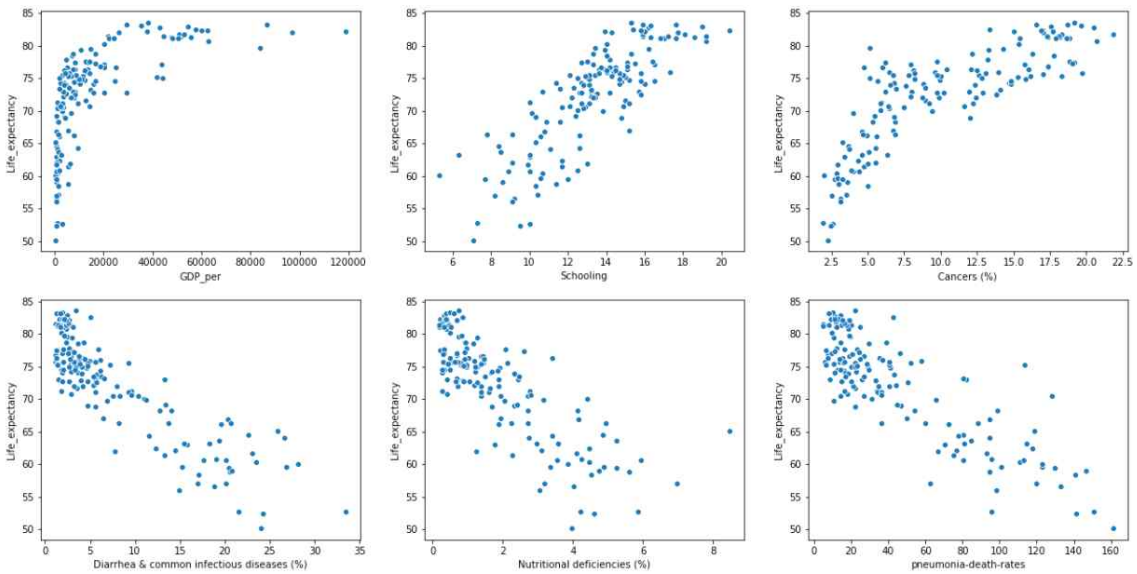
중간과제에서 서술했듯이 기대수명의 개념은 각 나이별 사망률을 구하여 0세의 아이가 사망할 것으로 기대되는 나이이다. 통계조사가 잘 이루어지고 있는 선진국들과는 관계가 없겠지만 그렇지 않은 국가들의 경우 몇 세부터 몇 세까지 나이의 사망률은 빠져있고 데이터의 결측치가 많은등 데이터에 오류가 있을 수 있다고 생각한다. 그럴 경우 이 모델을 사용하여 5개의 변수만으로 기대수명을 예측할 수 있다는 점에서 활용성도 있다고 생각하였다.

<부록 1> 데이터들의 출처

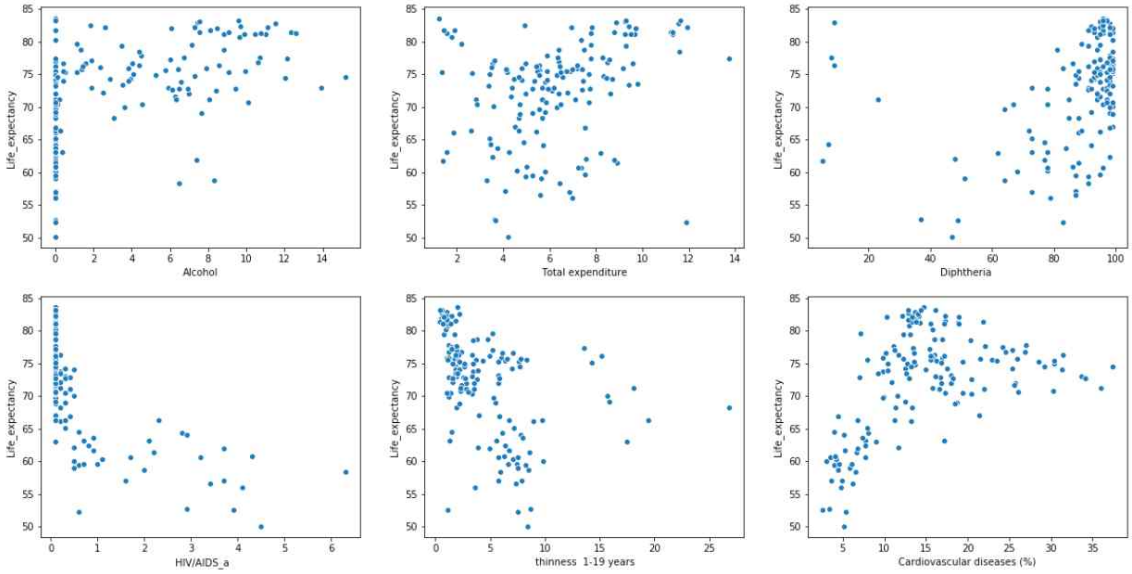
BMI :
https://apps.who.int/gho/data/node.main.A900A?lang=en
Alcohol:
https://www.who.int/data/gho/indicator-metadata-registry/imr-details/462
새로바꾼 기대수명:
https://databank.worldbank.org/reports.aspx?source=2&series=SP.DYN.LE00.IN&country=#
새로 바꾼 1인당 GDP: (통계청자료에서 GDP와 인구 데이터를 얻은뒤 GDP/인구 값으로 계산)
https://kostat.go.kr/wnsearch/search.jsp
health indicator:
https://www.kaggle.com/nxpnsv/country-health-indicators
각종 질병들의 DALYS
https://ourworldindata.org/grapher/pneumonia-death-rates-age-standardized

<부록 2> 설명변수 후보들과 기대수명과의 산점도

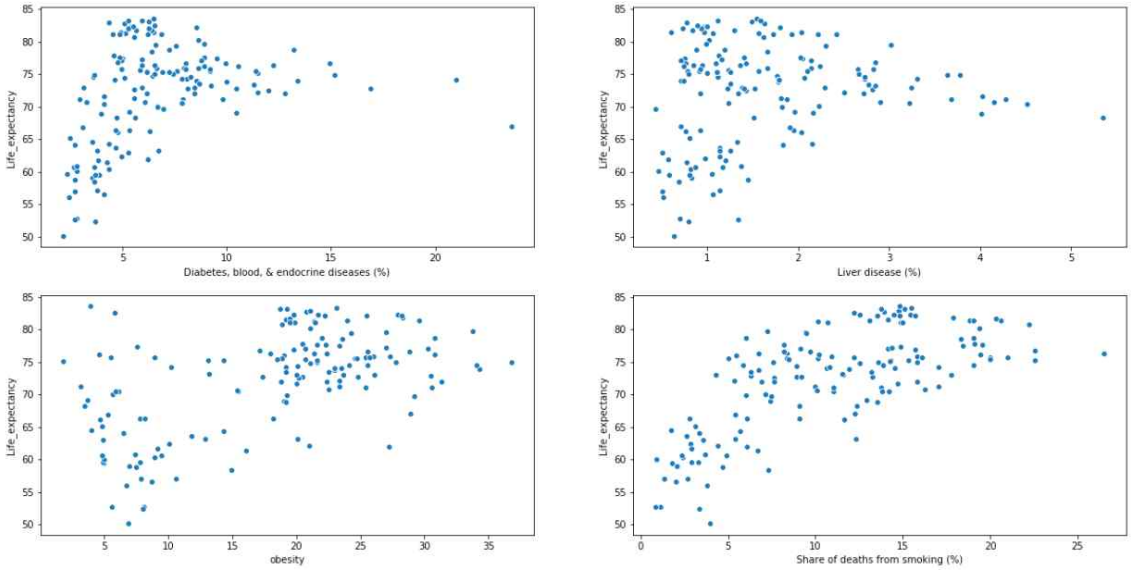
(왼쪽위부터 오른쪽으로) **1인당GDP , Schooling , Cancers , Diaherra & common diseases , nutritional deficiencies , pneumonia death rates**



Alcohol , Total_expenditure , Diphtheria , HIV/AIDS , Thiness1-19 , Cardivascular disease.

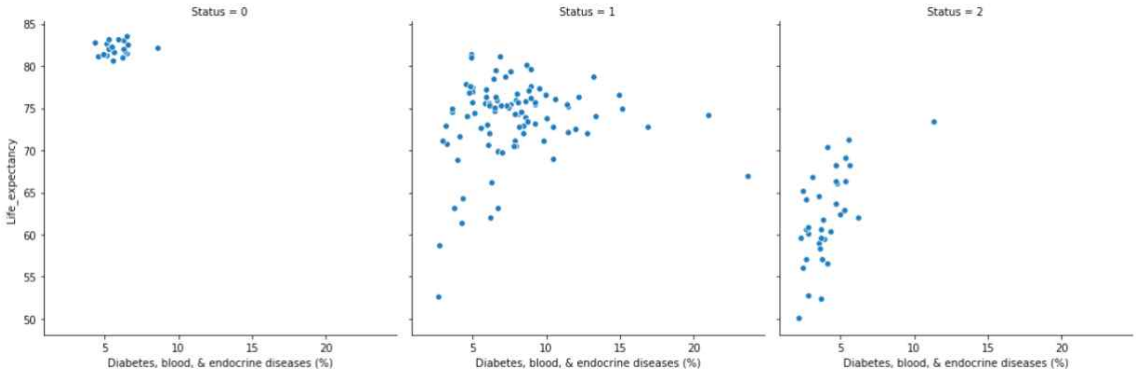


Diabetes , Liver disease , Obesity , Smoking

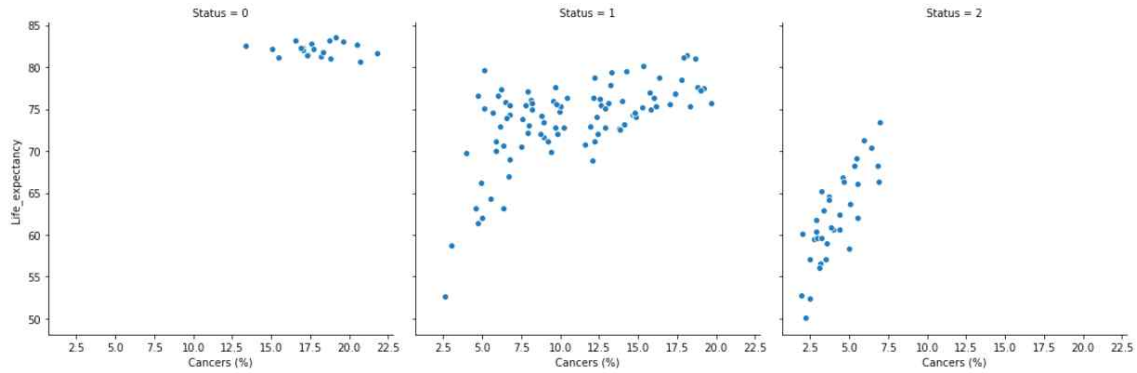


<부록 3> 각 수준별 (0= 선진국 , 1=중진국 , 2= 후진국) 변수와 기대수명 간의 산점도

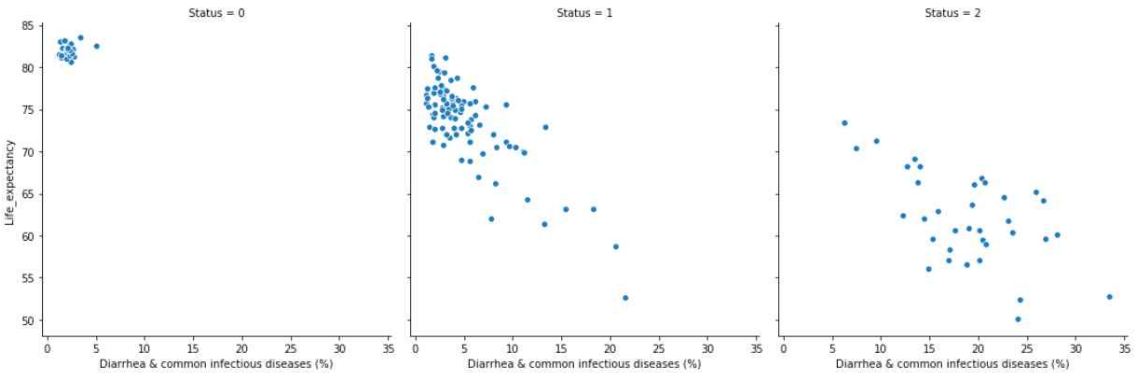
Diabete



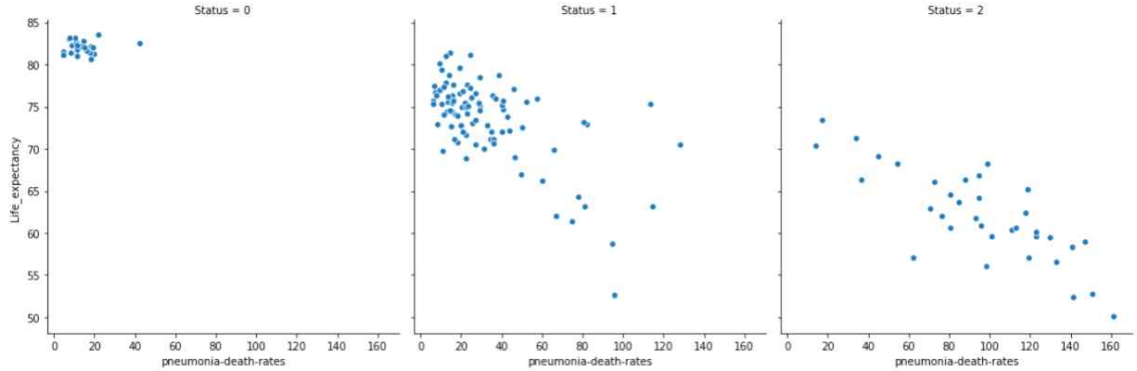
Cancers



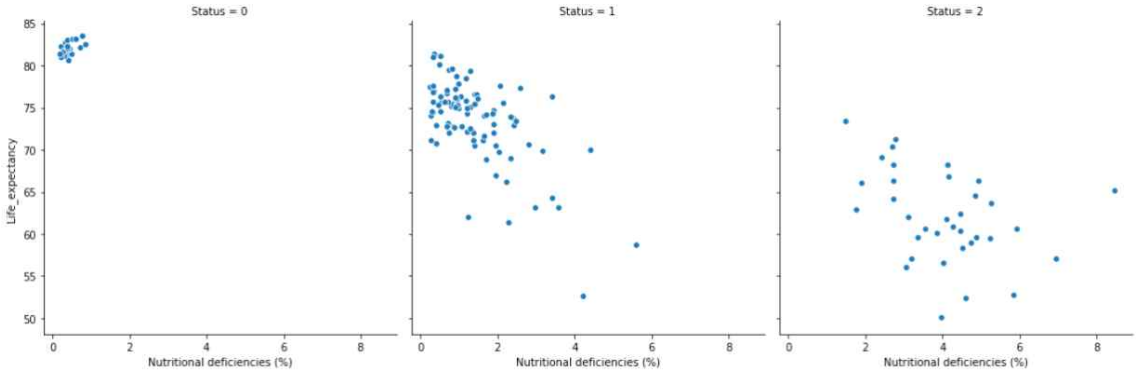
Diarrhea



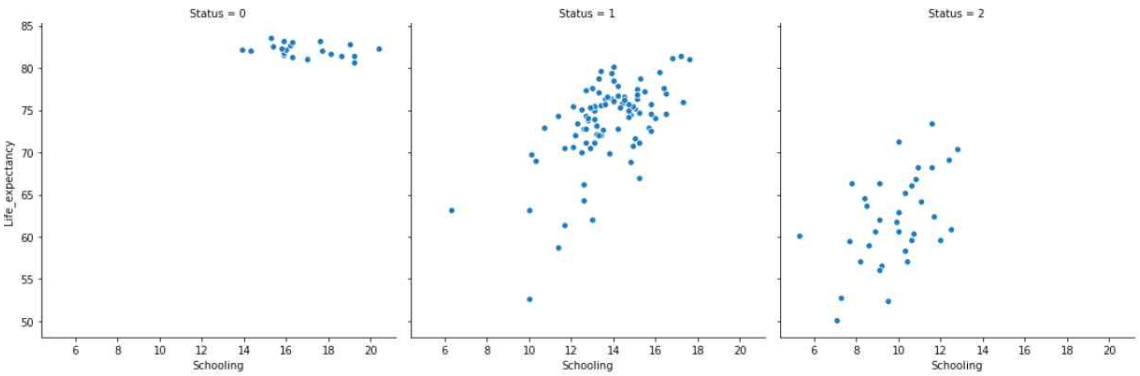
Pneumonia



Nutrition



Schooling

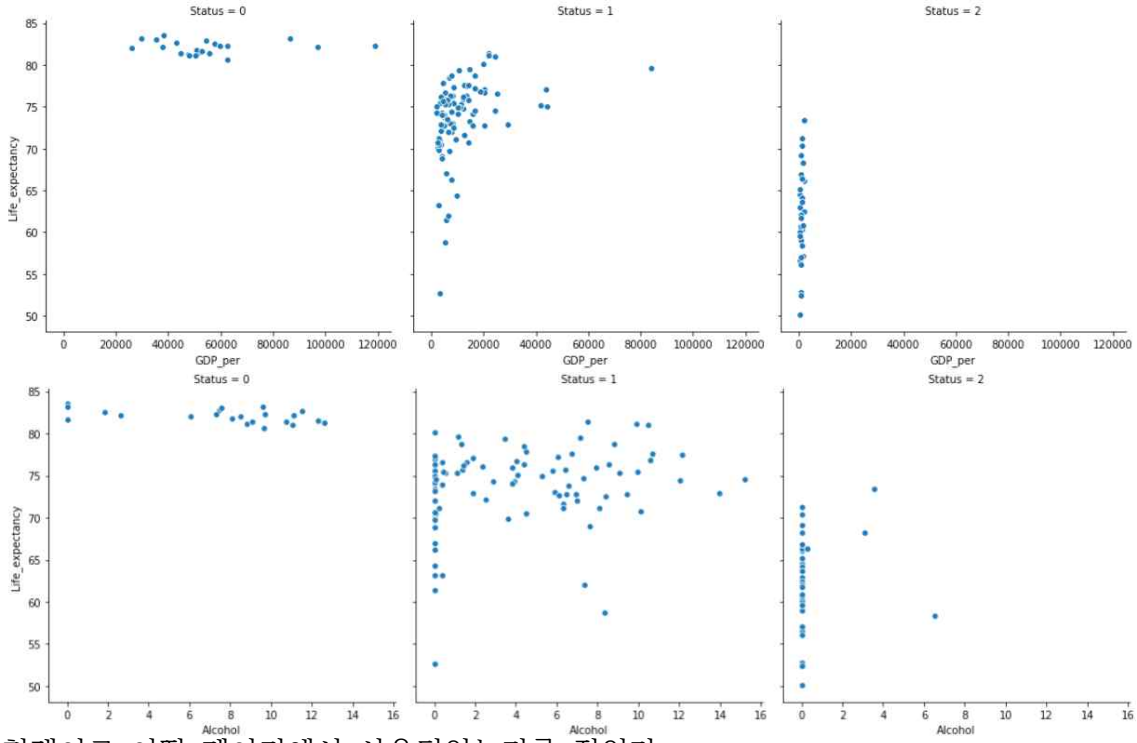


GDP_per

Alcohol

<부록 4> 사용 코드

모든 코드를 그대로 붙여 올리기보다는 각 단계별 사용했던 코드 중 이해에 필요한 만큼 올리는 것이 낫다고 생각하였다. 우선 산점도의 개수가 너무 많았고, 회귀모델을 적합하는 과정도 모델의 수가 꽤나 많기 때문에 그대로 올리기보다는 각 코드가 어떤



형태이고 어떤 페이지에서 사용되었는지를 적었다.

EDA- 파이썬의 Seaborn , matplotlib 사용

1. 각 수준별 분포도
Seaborn Library의 Distplot 사용 ,page6의 그림 3

```
f,ax=plt.subplots(1,3,figsize=(20,10))
sns.distplot(raw[raw['Status']==0]['Life_expectancy '],ax=ax[0])
ax[0].set_title('Developed Country')
sns.distplot(raw[raw['Status']==1]['Life_expectancy '],ax=ax[1])
ax[1].set_title('Middle Country')
sns.distplot(raw[raw['Status']==2]['Life_expectancy '],ax=ax[2])
ax[2].set_title('Developing Country')
plt.show()
```

2. 각 변수별 기대수명 산점도 , page 7 , <부록 2>
seaborn 라이브러리의 scatterplot 사용

```
f,ax=plt.subplots(2,3,figsize=(20,10))
sns.scatterplot(x="GDP_per", y="Life_expectancy ", data=raw,ax=ax[0,0])
plt.title='GDP_per_capita'
```

```
sns.scatterplot(x="Schooling", y="Life_expectancy ", data=raw,ax=ax[0,1])
plt.title='Schooling'
sns.scatterplot(x="Cancers (%)", y="Life_expectancy ", data=raw,ax=ax[0,2])
plt.title='Cancers (%)'
sns.scatterplot(x="Diarrhea & common infectious diseases (%)", y="Life_expectancy ", data=raw,ax=ax[1,0])
plt.title='Diarrhea & common infectious diseases (%)'
sns.scatterplot(x="Nutritional deficiencies (%)", y="Life_expectancy ", data=raw,ax=ax[1,1])
plt.title='Nutritional deficiencies (%)'
sns.scatterplot(x="pneumonia-death-rates", y="Life_expectancy ", data=raw,ax=ax[1,2])
plt.title='pneumonia-death-rates'
plt.show()
```

3. 국가수준별 변수들과 기대수명의 산점도 사용코드 ,page 8 , <부록 3>

seaborn 라이브러리의 relplot 사용

```
sns.relplot('Diabetes, blood, & endocrine diseases (%)','Life_expectancy ',data=raw,col='Status')
```

R

```
raw4=read.csv('raw3_6_18.csv')
```

가장 처음에 적합한 모델

```
lm1=lm(Life_expectancy~GDP_per+Schooling,data=raw4)
summary(lm1)
AIC(lm1)
```

수준별 적합 모델 그림 <13>

```
lm2=lm(Life_expectancy~GDP_per+GDP_per*factor(Developed_country)+GDP_per*factor(Middle_country)+Schooling,data=raw4)
summary(lm2)
AIC(lm2)
vif(lm2)
```

수준별 적합 모델 <그림 15)

```
lm7=lm(Life_expectancy~GDP_per+Schooling+Schooling*factor(Developed_country)+Schooling*factor(Developed_country),data=raw4)
summary(lm7)
vif(lm7)
AIC(lm7)
```

그림 <16>

```
lm9=lm(Life_expectancy~sqrt(GDP_per)+Schooling,data=raw4)
summary(lm9)
AIC(lm9)
```

그림 <17>

```
lm11=lm(Life_expectancy~sqrt(GDP_per)+Schooling+Cancers,data=raw4)
summary(lm11)
AIC(lm11)
```

그림 <18>

```
lm12=lm(Life_expectancy~sqrt(GDP_per)+Schooling+pneumonia+Nutrition,data=raw4)
summary(lm12)
AIC(lm12)
vif(lm12)

lm13=lm(Life_expectancy~sqrt(GDP_per)+Schooling+pneumonia+Cancers,data=raw4)
summary(lm13)
AIC(lm13)
```

그림 <19>

```
lm14=lm(Life_expectancy~sqrt(GDP_per)+Schooling+pneumonia+Alcohol,data=raw4)
summary(lm14)
AIC(lm14)
```

최종모델

```
lm15=lm(Life_expectancy~sqrt(GDP_per)+Schooling+pneumonia+Alcohol+Nutriti
on,data=raw4)
summary(lm15)
```

 # lm15 와 lm19는 같은 모델임

정규성 검정

```
shapiro.test(lm19$residuals)
```

독립성 검정

```
durbinWatsonTest(lm19)
```

변수들 scale화 하여 만든 모델

```
raw5=raw4
raw5$GDP_per= sqrt(raw5$GDP_per)
raw5=lapply(raw5,scale)
lm17=lm(Life_expectancy~GDP_per+Schooling+pneumonia+Alcohol+Nutrition,da
ta=raw5)
summary(lm17)
```

Relative weight analysis

```
library(ggplot2)
```
먼저 함수를 정의하고
```
relweights <- function(fit,...){
  R <- cor(fit$model)
  nvar <- ncol(R)
  rxx <- R[2:nvar, 2:nvar]
  rxy <- R[2:nvar, 1]
  svd <- eigen(rxx)
  evec <- svd$vectors
  ev <- svd$values
  delta <- diag(sqrt(ev))
  lambda <- evec %*% delta %*% t(evec)
```

```
  lambdasq <- lambda ^ 2
  beta <- solve(lambda) %*% rxy
  rsquare <- colSums(beta ^ 2)
  rawwgt <- lambdasq %*% beta ^ 2
  import <- (rawwgt / rsquare) * 100
  import <- as.data.frame(import)
  row.names(import) <- names(fit$model[2:nvar])
  names(import) <- "Weights"
  import <- import[order(import),1, drop=FALSE]
  dotchart(import$Weights, labels=row.names(import),
           xlab="% of R-Square", pch=19,
           main="Relative Importance of Predictor Variables",
           sub=paste("Total R-Square=", round(rsquare, digits=3)),
           ...)
  return(import)
}
```

함수정의2

```
plotRelWeights=function(fit){
  data <- relweights(fit)
  data$Predictors <- rownames(data)
  p <- ggplot(data=
data,aes(x=reorder(Predictors,Weights),y=Weights,fill=Predictors))+geom_bar(st
at="identity",width=0.5)+ggtitle('Relative Importance of predictor
Variables')+ylab(paste0("% of R-square \n(total R-Square=
",attr(data,"R-square"),")"))+geom_text(aes(y=Weights-0.1,label=paste(round(Weig
hts,1),"%")),hjust=1)+guides(fill=FALSE)+coord_flip()
  p
}
```

적합한 회귀모형을 함수에 넣는다

```
plotRelWeights(lm15)

new_country=read.csv('test_set.csv')
predict(lm19,newdata=new_country,interval='prediction')
summary(lm19)
```

테스트 국가들 기대수명 예측

```
new_country=read.csv('test_set.csv')
new_country$GDP_per=as.numeric(new_country$GDP_per)
predict(lm19,newdata=new_country,interval='prediction')
```