R을 이용한 다중 선형 회귀 분석 :
사회적 요인이 기대수명에 미치는 영향 Modeling

2017. 12. 03.

학과·학년	통계학과 2학년
학번	2016150460 2016150422 2016150455
제출자	심정은 양수형 이혜진

목차

I. 서론

1. 연구 목적, 방향 및 연구 대상

우리나라 출생아의 기대수명은 1960년에서 2014년까지 급속하게 증가했으며, OECD 통계에 따르면, 1960년 이후 우리나라는 OECD 국가 중에서 기대수명이 가장 빠르게 증가한 것으로 나타났다. 우리나라뿐만 아니라 기대수명 증가는 전 세계에서 공통적으로 나타나는 현상이다.[1] 이러한 기대수명 증가에 영향을 미치는 사회적 요인으로 언급되는 것은 크게 경제 성장, 의료기술의 발전 등이 있다. 기대수명에 영향을 미칠 것으로 보이는 요인들 중에서 실제로 기대수명과의 연관성을 알아보기 위해 다중 선형 회귀 분석으로 이에 대해 탐구해보았다.

본 연구는 다중선형회귀모델을 설정하여 다중회귀분석을 해보고, 더 적합한 모델을 선택해보도록 한다. 범용성과 전문성을 갖춘 R 소프트웨어와 lmtest, leaps, corrplot, DAAG 패키지를 분석에 이용하였다.

2. 모형 및 변수 설명

본 연구에서는 R 소프트웨어를 이용하여 인간의 기대수명(life expectancy) 변화를 설명하는 회귀모형 설정하고, 각 변수마다 최신 년도를 기준으로 전 세계 248개 국가에서 수집한 데이터를 전처리 과정을 통해 147개 국가로 축소하여 사용한다. 구체적인 회귀모형은 다음과 같다.

$$life_i = \beta_0 + \beta_1 gdp_i + \beta_2 sani_i + \beta_3 pre_i + \beta_4 pri_i + \beta_5 \sec_i + \beta_6 ter_i + \beta_7 smo_i + \beta_8 ob_i + \beta_9 al_i + \beta_{10} co2_i + \beta_{11} hiv_i$$

국가별 기대수명을 Response variable로 두고, 여러 사회적 환경을 대변하는 지수를 Explanatory variable로 두었다.

여기서, response variable과 explanatory variable의 기준은 다음과 같다.

response variable :

A. life : 기대수명(Life expectancy at birth, total(year)) / 2015년[2]

explanatory variable :

B. gdp : GDP per capita (current US$) / 2016년[3]

C. sani : Improved sanitation facilities (% of population with access) / 2015년[4]

D. pre : Gross enrollment ratio, pre-primary, both sex (%) / 2015년[5]

E. pri : Gross enrollment ratio, primary, both sex (%) / 2015년[6]

F. sec : Gross enrollment ratio. secondary, both sex (%) / 2015년[7]

G. ter : Gross enrollment ratio. tertiary, both sex (%) / 2015년[8]

H. smo : smoking 2013 daily cigarette, both sexed, aged-standardized rate / 2013년[9]

I. ob : prevalence of obesity, BMI 〉= 25, 18+, age-standardized estimate / 2016년[10]

---

1) 최용옥, 급속한 기대수명 증가의 함의(Longevity Risk in Korea), KDI FOCUS, NO.69, (Korea Development Institute), 2016, p3
2) https://data.worldbank.org/indicator/SP.DYN.LE00.IN
3) https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
4) https://data.worldbank.org/indicator/SH.STA.ACSN?name_desc=false
5) https://data.worldbank.org/indicator/SE.PRE.ENRR
6) https://data.worldbank.org/indicator/SE.PRM.ENRR
7) https://data.worldbank.org/indicator/SE.SEC.ENRR
8) https://data.worldbank.org/indicator/SE.TER.ENRR
9) http://apps.who.int/gho/data/view.main.TOB30011
10) http://apps.who.int/gho/data/view.main.CTRY2430A

J. al : Total alcohol consumption per capita
(liters of pure alcohol, projected estimates, 15+ years of age) / 2015년[11)]
K. co2 : CO2 emissions (metric tons per capita) / 2014년[12)]
L. hiv : Prevalence of HIV, total (% of population ages 15-49) / 2016년[13)]

II. 본론

1. 전처리

각 변수를 다른 데이터로부터 수집하다 보니, 변수마다 수집된 년도가 달랐고, 무엇보다 데이터 중에서 NA가 많았다. 이를 해결하기 위해, 처음에는 수집된 데이터에서 2013년~2016년도 자료를 추출한 뒤, 각 국가별로 최신년도의 자료를 사용하도록 하였다. 만약, 2013년~2016년 자료가 다 존재하지 않는다면, NA로 처리하였고, 그런 국가는 분석에서 제외했다. 그 후, 공통된 국가 기준으로 inner join을 하여 데이터를 합쳐(merge)해 주었더니, 248개의 세계 국가 중 observation인 국가가 60개밖에 남지 않았다. 60개의 observation은 분석에 있어서 그 수가 적다고 판단하여, KNN을 사용하여 데이터 전처리 과정을 다시 거쳤다.

1) NA 처리 방법

· 처음 NA 처리

우선, 이 보고서의 목적은 과거 유사한 목적의 논문의 데이터가 오래되어 최신의 자료로 대체함과 동시에 분위수 회귀 대신 다중 회귀 분석으로 분석해 보는 것이다. 그러기 위해서 처음에 모인 데이터에서 2013년 이후의 데이터만 선별하였다. 뿐만 아니라, 참고 논문에서는 백인, 황인, 흑인 3인종의 dummy variable을 사용했지만, 다인종 국가의 경우 규정이 어려울 것 같아 Asia, America, Africa, Europe 4개 대륙으로 분류하여 dummy variable을 넣었다.

데이터 처리에 있어서 무엇보다도 중요한 것은 na를 처리하는 과정이다. na가 있다면 일반적 함수 계산이 불가하기 때문에 na를 어떻게 처리할지에 관해 가장 먼저 논의해 보았다. na 값을 0으로 처리하는 방법과 na를 제거하는 방법 중 처음에 선택한 방법은 na를 지우는 방법이었다.

그를 위해서 각 life, gdp, sanitation, pre-primary, primary, secondary, tertiary, smoking(by cigarette), obesity, alcohol, Co2 emissions과 hiv 각 변수 데이터에서 na가 있는 row는(observation인 국가) 전부 제거하였다.

---

11) https://data.worldbank.org/indicator/SH.ALC.PCAP.LI
12) https://data.worldbank.org/indicator/EN.ATM.CO2E.PC
13) https://data.worldbank.org/indicator/EN.ATM.CO2E.PC (1990~2016년) 데이터에 제외되어 있는 주요 국가를 같은 기준으로 조사된 자료 https://data.humdata.org/dataset/prevalence_of_hiv_total_of_population_aged_15-49/resource/c5f56338-471b-4aaf-b5b9-b1f7db160bc1 (1950~2014년)를 이용해서 업데이트 해주었다.

```
recentYear <- function(data) {
  data1 <- data[,c(1,2)]
  data <- data[,-c(1,2)]
  for(i in 1:nrow(data)){
    index <- !is.na(data[i,])
    index.1 <- order(index)[length(index)]
    data1$referenceYear[i] <- colnames(data)[index.1]
    data1$total[i] <- data[i, index.1]
  }
  return(data1)
}
```

이후, recentYear 함수를 통해 각 row인 국가별로 최신의 year을 뽑기로 했다. 다만 이런 방식은, 각 국가별로 최신 자료의 년도가 다르기 때문에 분석하였을 때의 결과의 신뢰성에 다소 문제가 있을 것 같은 문제가 발생했다. 이렇게 각 변수의 국가별 recentYear로 정리된 파일을 data frame으로 합쳐보니 total observation 개수가 60개가 되었다. 그와 dummy variable를 합쳐서 Total 데이터 셋을 만들어 주었다.

Data	
al	240 obs. of 3 variables
co2	251 obs. of 3 variables
cont	248 obs. of 3 variables
df	60 obs. of 18 variables
dummy	248 obs. of 6 variables
edu	162 obs. of 6 variables
gdp	246 obs. of 3 variables
hiv	162 obs. of 3 variables
ob	182 obs. of 3 variables
sani	240 obs. of 3 variables
smo	124 obs. of 3 variables
Total	60 obs. of 18 variables

그렇게 정리된 전체 data frame의 앞 13개의 observation이다.

	Country.Name	Country.Code	aa	af	am	eu	GDP	Sanitation	pre.primary	primary	secondary	tertiary	Daily.smoking	obesity	Alchol	CO2	HIV
1	Albania	ALB	0	0	0	1	4146.8962	93.2	88.60224	111.87708	95.76549	8.66280	23.8	57.7	93.2	1.97876331	0.1
2	Argentina	ARG	0	0	1	0	12449.2169	96.4	72.75394	116.34584	88.50236	28.17496	18.1	62.7	96.4	4.74679746	0.4
3	Armenia	ARM	1	0	0	0	3606.1521	89.5	52.39516	109.98889	102.70546	82.91739	24.7	54.4	89.5	1.90275891	0.2
4	Australia	AUS	1	0	0	0	49927.8195	100.0	124.91998	97.05287	100.02189	90.30650	13.7	64.5	100.0	15.39859985	0.1
5	Azerbaijan	AZE	1	0	0	0	3876.9364	89.3	23.87579	102.97016	166.80847	25.48320	18.6	53.6	89.3	3.93156061	0.1
6	Bahrain	BHR	1	0	0	0	22354.1671	99.2	55.85834	97.21818	80.79221	43.26323	27.1	65.8	99.2	23.44975483	0.1
7	Bangladesh	BGD	1	0	0	0	1358.7798	60.6	31.22427	87.99263	102.13090	13.44080	19.6	20.0	60.6	0.45914196	0.1
8	Barbados	BRB	0	0	1	0	16096.8926	96.2	84.20607	115.34079	84.19769	10.92693	5.0	52.4	96.2	4.49017767	1.3
9	Belarus	BLR	0	0	0	1	4989.2546	94.3	103.22911	101.17117	71.54241	87.94074	22.6	59.4	94.3	6.70195770	0.4
10	Benin	BEN	0	1	0	0	789.4404	19.7	23.90071	104.19298	63.52453	15.36278	7.4	29.5	19.7	0.61421385	1.0
11	Brazil	BRA	0	0	1	0	8649.9485	82.8	92.17724	97.13804	96.07947	30.84478	12.7	56.5	82.8	2.59438828	0.6
12	Bulgaria	BGR	0	0	0	1	7350.7958	86.0	82.91228	120.43306	107.11942	73.93420	30.0	61.7	86.0	5.87161587	0.1
13	Burkina Faso	BFA	0	1	0	0	649.7305	19.7	4.14232	128.98335	99.01635	5.56218	16.2	23.2	19.7	0.16201881	0.8

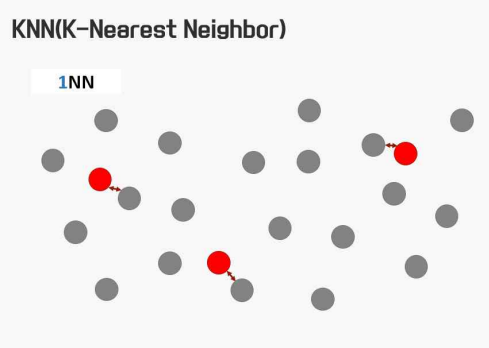
하지만, observation이 60개 밖에 남지 않아 데이터 분석에 다소 어려움이 다소 있었다. 뿐만 아니라, 위에서 언급했듯, 한 변수 내에서도 국가별로 recentYear가 달라 분석의 신뢰성이 떨어질 수 있다고 판단해 다른 방법으로 전체 data set을 구하기로 변경하였다. 또한, 대륙별 dummy variable까지 함께 다중 회귀 분석하기에는 변수가 너무 많아 대륙별 변수를 제거하고 나머지 11개의 explanatory variable과 한 개의 response variable로만 분석을 진행하였다.

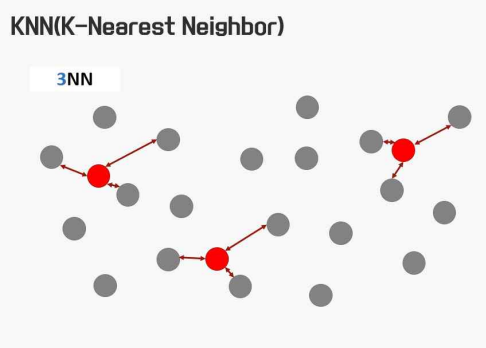
· KNN(K-Nearest Neighbor) 방법 사용

많은 NA를 처리하기 위해 KNN 방법을 통해서 NA를 처리하였다.

KNN(K-Nearest Neighbor) 알고리즘은 기존 데이터 중 가장 유사한 K개의 데이터를 이용해서 새로운 데이터를 예측하는 방법이다. 수치형 데이터인 경우, 새로운 데이터와 가장 가까운 k개의 데이터의 평균의 평균을 이용한다. 명목형 데이터인 경우, K개의 데이터 중 많이 나온 분류 항목을 선택하거나 혹은 가중치를 주어 선택하도록 한다.[14)]

필자는 NA를 KNN을 사용해서 나온 값으로 채우고자 한다. 빨간색 원을 NA라고 가정한다.





K가 1일 때, 1NN이라고 하며, NA(빨간색 점)와 가장 가까운 1개의 데이터의 수치를 따라간다. K가 3이라고 할 때, 3NN이라고 하며, NA(빨간색 점)과 가장 가까운 3개의 데이터 수치의 평균이 NA의 값과 유사하다고 판단하여 그 값을 이용한다.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$ [15)]

수치형 데이터에서 가장 흔하게 사용되는 거리 척도는 유클리드 거리이다. 따라서 가장 가까운 데이터를 찾는데 유클리드 거리를 사용하였다.

유클리드 거리를 사용하여 필자는 k=5인 경우, 즉 5NN을 사용하여, NA를 처리하였다.

5NN 방법을 위해 다음과 같은 코드를 작성하였다.

14) http://www.datamarket.kr/xe/board_BoGi29/9880
15) http://www.saedsayad.com/k_nearest_neighbors_reg.htm

```r
KNN <- function(data, year){ #data : raw data, year : 기준년도 # column
  if(length(which(is.na(data[,year]))) == 0) return(data[,c(1,2, year)])
  na.row <- which(is.na(data[, year])) #기준 년도에서 NA인 row추출

  for(i in 1:length(na.row)){
    col <- !is.na(data[na.row[i],]) #key observation에서 NA가 아닌 col 추출
    collected.col <- data[, col] #data에서 col column들을 추출
    key <- collected.col[na.row[i],] #key observation의 row 추출
    index <- complete.cases(collected.col) #NA가 하나도 없는 row 추출
    non.na <- collected.col[index,] #NA를 제거함.
    d <- apply(as.data.frame(non.na[,-c(1,2)]), 1, "-", key[,-c(1,2)]) #두 값의 차
    d2 <- unlist(d) ^ 2 #차의 제곱
    d2.matrix <- as.data.frame(matrix(d2, length(d2)/(length(collected.col)-2),
                               length(collected.col)-2, byrow = T)) #vector b를 data.frame으로 형변환
    colnames(d2.matrix) <- colnames(non.na)[-c(1, 2)]
    p.length <- apply(c, 1, sum) #각 년도별로 구해진 길이들의 합
    o.p <- order(p.length) #길이들의 합을 최소 순서로 배열
    n.point <- o.p[which(!is.na(data[o.p, year]))][1:5] #year 칼럼에 NA가 아닌 것들 중에서 작은거 5개
    n.data <- data[n.point, year] #key와 가까운 5개의 point의 점수 추출
    data[na.row[i], year] <- mean(n.data, na.rm = T) # 추출된 근처 값들의 평균
  }
  return(data[,c(1, 2, year)])
}
```

위의 수치형 KNN method를 사용자 지정 함수로 만들어 보았다. 우선 분석에 사용할 기준 년도를 결정한다. (ex 2015) 그 후 그 해의 observation 중 na가 있는 국가를 뽑아낸다. 이제 2015년에 na가 있어 뽑힌 국가 중 첫 번째 국가의 역대 년 기록 중, na가 있는 열(년도)을 다시 제거한다. 그렇게 하면 기준년도에 na가 있는 행(na.row) 즉 기준년도에 na를 포함한 여러 국가들의 기록은 na가 없는 년도의 기록만 남을 것이다. 그렇다면 해당 국가의 남아있는 년도를 기준으로 다른 모든 248개의 국가들도 그 년도만을 남겨 na.row 첫 국가의 수치값과 가장 유사한 5개의 국가를 선택한다. 이 국가들을 선택하는 방법은 유클리디안 거리를 사용한다. 5개의 국가를 선택하게 되면 이제 2015년도로 돌아가 선택된 5개 국가의 평균값을 계산해 2015년 값이 na던 최초의 na.row 국가에 평균값을 넣는다. 그렇게 na.row에 속한 모든 국가에 대해 반복하여 2015년 na 값을 knn method를 사용하여 값과 근처에 있는 국가들 5개의 평균값으로 대체한다. 만약 5개의 국가를 뽑아 2015년의 그 국가들의 값을 보았으나 na인 국가가 있다면, 6개째의 국가를 뽑아 그로 대체한다.

2) 데이터 셋 만들기

5NN 방법을 적용한 뒤, 다중선형회귀분석을 위한 데이터 셋을 다음과 같이 처리하였다.

```
> dim(df)
[1] 147  14
> head(df)
  Country.Name Country.Code     life          gdp  sani       pre       pri      sec      ter    smo   ob   al       co2 hiv
1  Afghanistan          AFG 63.29820 1.946902e+10  31.9  74.26562 111.87708 55.64441 56.21012 15.264 23.0  1.0  0.299445 0.1
2       Angola          AGO 61.18934 8.963316e+10  51.6  68.32977 105.32842 79.80781  9.30802 15.264 27.5  7.6  1.291328 1.9
3      Albania          ALB 78.20315 1.192689e+10  93.2  88.60224 113.69980 95.76549 58.10995 21.300 57.7  6.6  1.978763 0.1
4    Argentina          ARG 76.29302 5.460000e+11  96.4  60.21817 103.77530 88.24808 52.54843 17.300 62.7  7.6  4.746797 0.4
5      Armenia          ARM 74.20620 1.054733e+10  89.5  52.39516  98.46662 88.50236 44.30950 22.700 54.4  5.5  1.902759 0.2
6    Australia          AUS 82.45122 1.200000e+12 100.0 124.91998 102.20782 98.71488 50.67717 11.000 64.5 12.6 15.398600 0.1
```

위에 소개된 일련의 과정을 통해서 나온 데이터 셋은 248개 국가 중에서 147개의 국가로 관측치가 백여 개의 관측치가 제거 되었다.

2. 데이터 분석

1) 다중 회귀 모형

```
> round(obj$coefficients, 4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.2283     3.3285 13.8888   0.0000
gdp.1         0.0000     0.0000  2.8315   0.0054
sani.1        0.1867     0.0158 11.8429   0.0000
pre.1         0.0326     0.0125  2.6137   0.0100
pri.1         0.0517     0.0277  1.8652   0.0644
sec.1         0.0217     0.0174  1.2445   0.2155
ter.1         0.0385     0.0178  2.1671   0.0320
smo.1        -0.1100     0.0574 -1.9160   0.0575
ob.1          0.0432     0.0211  2.0499   0.0424
al.1          0.0644     0.0866  0.7437   0.4583
co2.1         0.0177     0.0558  0.3175   0.7513
hiv.1        -0.3816     0.0775 -4.9273   0.0000
```

beta($\hat{\beta}$)의 회귀계수는 각각

$\widehat{\beta_0}= 46.2283, \widehat{\beta_1}= 0.0000, \widehat{\beta_2} = 0.1867, \widehat{\beta_3} = 0.0326, \widehat{\beta_4}= 0.0517, \widehat{\beta_5} = 0.0217, \widehat{\beta_6} = 0.0385,$
$\widehat{\beta_7}=-0.1100, \widehat{\beta_8}= 0.0432, \widehat{\beta_9}= 0.0644, \widehat{\beta_{10}}= 0.0177, \widehat{\beta_{11}}=-0.3816$

으로 추정할 수 있다.

$$\widehat{life}_i= 46.2283 + 0 \times gdp_i + 0.1867\, sani_i + 0.0326\, pre_i + 0.0517\, pri_i + 0.0217\, \mathrm{sec}_i + 0.0385\, ter_i - 0.1100\, smo_i + 0.0432\, ob_i + 0.0644\, al_i + 0.0177\, co2_i - 0.3816\, hiv_i$$

여기에서 gdp의 coefficient를 보면 소수점 4번째 자리까지 rounding 했기 때문에 0.0000으로 나온 것을 볼 수 있을 것이다. 이렇게만 보면 gdp가 life와 연관성이 전혀 없는 것으로 나타난다. 하지만, 선행 논문과 비교해 보았을 때 잘못된 결과임을 알 수 있다. 이는 gdp의 수치가 너무 커서 발생한 결과로, 따라서 log transformation을 수행한 모델도 구해보았다. 이는 아래 3) log transformation에서 더 자세히 확인해 볼 수 있다.

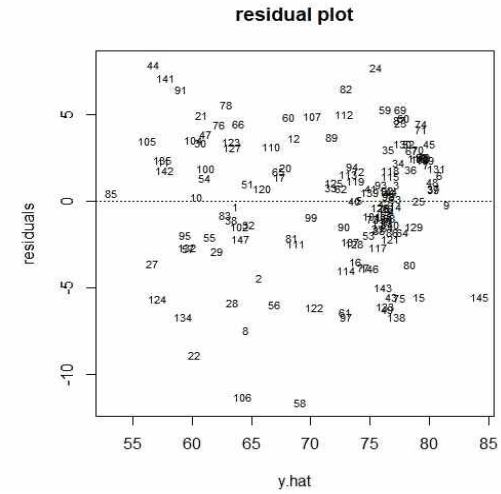
```
> summary(reg)

Call:
lm(formula = life ~ gdp + sani + pre + pri + sec + ter + smo +
    ob + al + co2 + hiv)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5956  -2.0054   0.4598   2.4927   7.8858

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.950e+01  3.245e+00  15.252  < 2e-16 ***
gdp          3.247e-13  1.761e-13   1.844   0.0674 .
sani         1.949e-01  1.636e-02  11.910  < 2e-16 ***
pre          3.285e-02  1.330e-02   2.470   0.0147 *
pri          2.877e-02  2.744e-02   1.048   0.2963
sec          1.803e-02  1.847e-02   0.976   0.3309
ter          3.271e-02  1.889e-02   1.731   0.0857 .
smo         -6.419e-02  6.053e-02  -1.061   0.2908
ob           1.642e-02  2.161e-02   0.760   0.4486
al           3.008e-02  9.179e-02   0.328   0.7436
co2          2.354e-02  5.964e-02   0.395   0.6937
hiv         -3.755e-01  8.252e-02  -4.550 1.18e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.845 on 135 degrees of freedom
Multiple R-squared:  0.7992,    Adjusted R-squared:  0.7829
F-statistic: 48.86 on 11 and 135 DF,  p-value: < 2.2e-16
```

이렇게 만들어진 모델을 summary 함수를 통해 더 자세히 보도록 하자. summary를 사용하여 자세한 estimation값과 각 variable 별 t-test 결과 또한 확인할 수 있다. $H_0 : \beta_j = 0 \;\; vs \;\; H_a : \beta_j \neq 0 \;\; (j = 1,2,\cdots,11)$하에서 가설검정을 하면, 각 variable의 T value를 구할 수 있다. 유의수준 0.05 수준에서 P-value를 살펴보면 sanitation과 pre-primary 그리고 hiv가 유의하다고 볼 수 있다. 모델 전체의 적합성을 살펴보는 F statistic이 48.86로 상당히 크고 P-value 또한 0과 거의 가까울 정도로 매우 작은 값이기 때문에 이 모델이 적합하다고 볼 수 있는 근거가 된다. 게다가 Adjusted R^2이 0.7829로 상당히 높아 모델의 설명력이 높다고 판단한다.

2) Assumption Test

· heteroscedasticity

Assumption 중 homoscedasticity(등분산성)가정을 확인하기 위해서는 x축이 fitted value, y축이 residual인 residual plot에서 경향성 없이 무작위하게 분포되어 있음을 보면 된다.

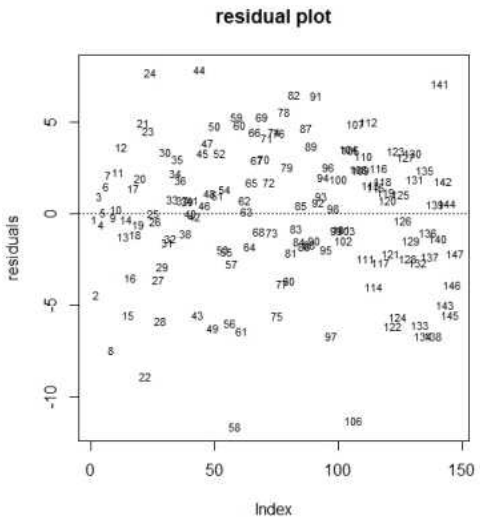


위의 residual plot을 육안으로 보았을 때 75~80 사이에 상당히 많은 값들이 몰려있는 것으로 보아 경향성이 다소 나타난다. 따라서 heteroscedasticity(이분산성)이 있는 것으로 보인다. 이 경우에는 Weighted least square(WLS) estimation 혹은 transformation을 사용해야 한다. 그래서 log를 취할 수 있을 것처럼 보이는 일부 변수에 transformation을 해 보기로 한다.

· independence

```
        Durbin-Watson test

data:  life ~ gdp + sani + pre + pri + sec + ter + smo + ob + al + co2 +
    hiv
DW = 2.2321, p-value = 0.9236
alternative hypothesis: true autocorrelation is greater than 0
```

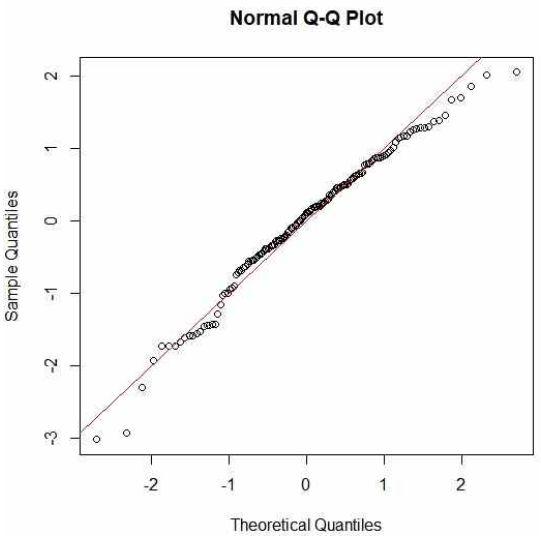
Durbin – Watson statistic은 residual의 autocorrelation을 측정하는 도구이다. 이 DW test 값이 충분히 작다면 H0를 기각할 수 있고, 너무 크다면 H0를 기각하는데 실패한다. 주어진 모델을 통해 independence assumption을 검정하기 위해 DW test를 실시한 결과, DW = 2.2321이고 그에 따른 P-value가 0.9236으로 매우 크기 때문에 independent하다고 볼 수 있다.



index residual plot이 무작위하게 분포되어있기에 independent함을 다시 한 번 확인할 수 있다.

· normality

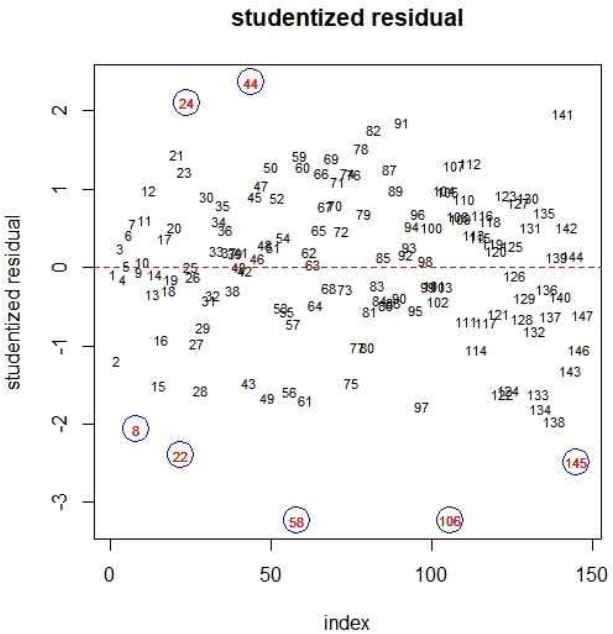
ordered standardized residual과 normal score을 비교했을 때, plot에서 볼 수 있듯이 거의 모든 점들이 y=x 위에 위치한다. 그러므로 정규성을 만족한다고 볼 수 있다. 그러나 n=147개로 충분히 크기 때문에 CLT에 의해 정규분포를 따르게 된다. 따라서 분석에 있어서 필수적인 assumption test는 아니다.



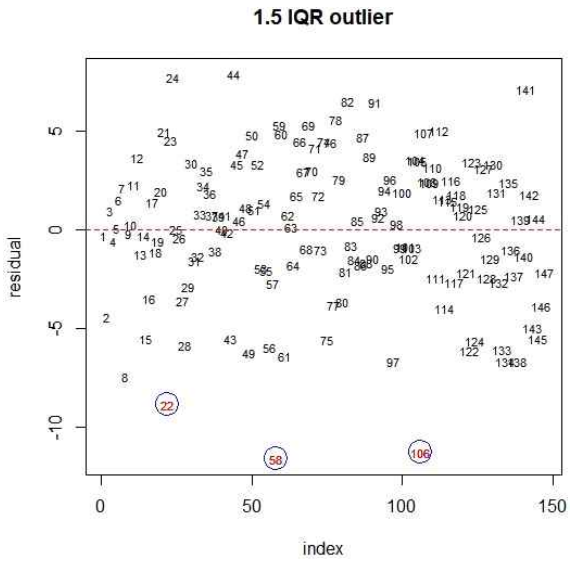
3) Outlier and Influential Point detection

· Outlier

outlier(이상치)는 주어진 회귀 모델에 의해 잘 설명되지 않는 데이터 점들을 뜻한다. Outlier detecting에서는 Studentized Residual을 사용해서 outlier를 찾아내는 방법이 일반적이다.



이를 보면, outlier의 번호와 개수를 알 수 있다. outlier는 총 7개로 판단할 수 있다.



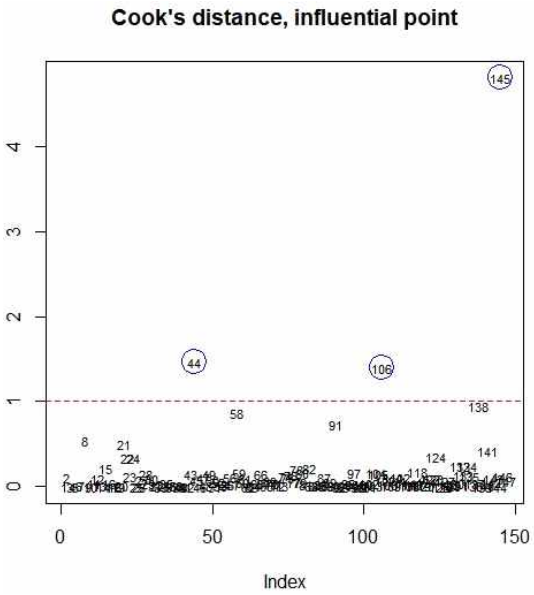
1.5 x IQR을 기준을 쓴다면, oulier은 다음과 같이 3개이다. 이 세 개의 outlier은 위의 점과 겹친다는 것을 알아차릴 수 있다.

・Influential point

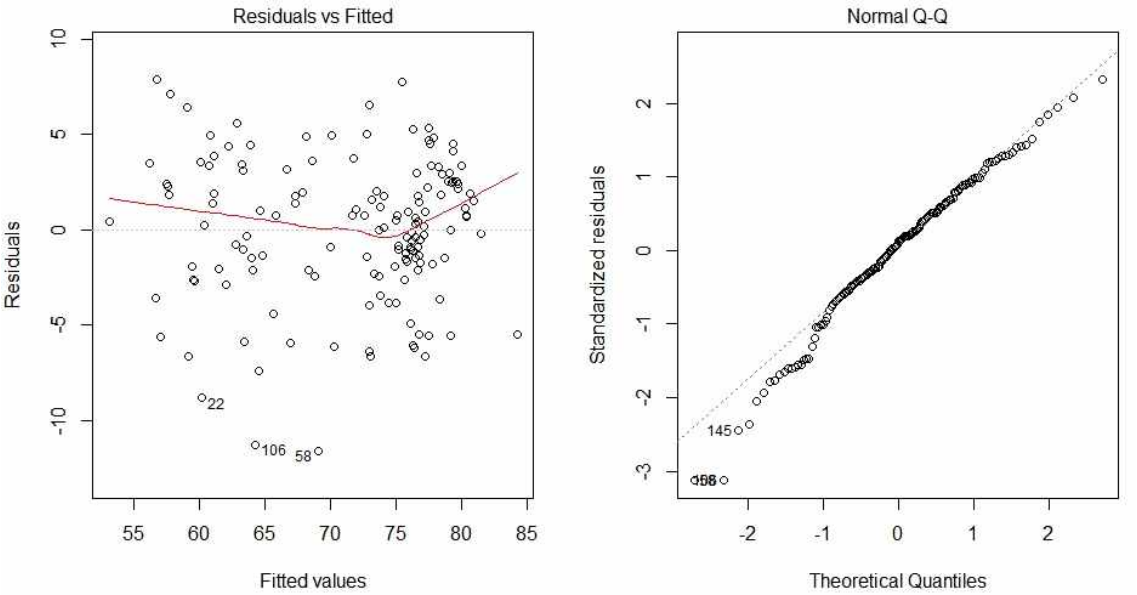
influence observation은 데이터 분석에 포함되었을 때, 큰 영향을 주는 observation이다. influence observation을 찾아내는 방법은 DFFITS, Cook's Distance가 있는데 필자는 Cook's Distance를 사용해 보았다.

Cook's Distance $C_i = \frac{1}{\widehat{\sigma^2}} \sum_{m=1}^{n} (\widehat{y_m} - \widehat{y_{m(i)}})^2 = (\frac{r_i^2}{p+1})(\frac{\widehat{h_{ii}}}{1-h_{ii}})$

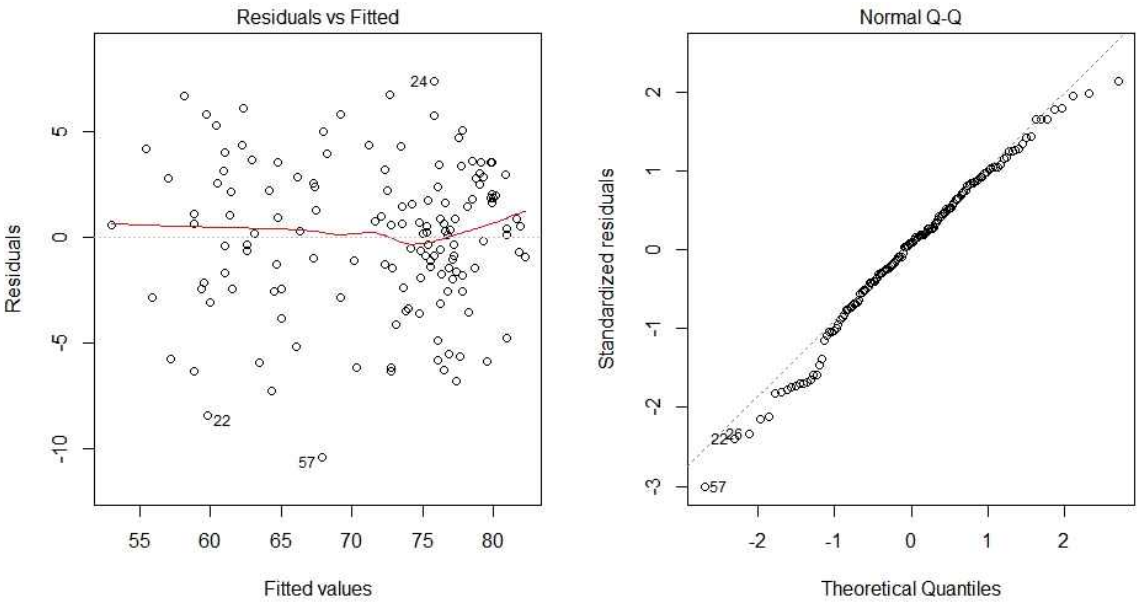
$\widehat{y_m}$은 full model을 의미하고 $\widehat{y_{m(i)}}$은 i 번째 observation을 제거한 후의 model이다. 식에서 인지할 수 있듯이 full model과 i 번째 observation을 제거한 후의 model의 fitting line 간의 거리 차이를 측정한 것으로 두 모델의 변화를 이용하여 influence observation을 찾아낸다. Cook's distance는 F 분포를 따르지만, 실전에서는 1보다 크면 influential하다고 판단한다.



아래의 plot을 통해서 44번째, 106번째, 145번째 observation의 Cook's distance가 1보다 크기 때문에 influential observation으로 간주된다.



〈influential observation이 제거되지 않은 assumption test〉

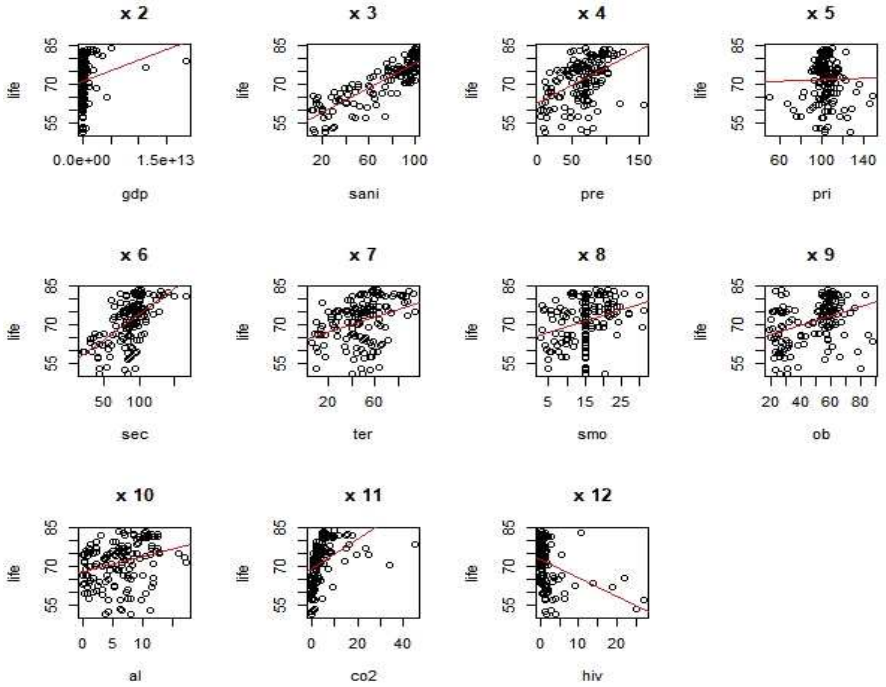


〈influential observation이 제거한 assumption test〉

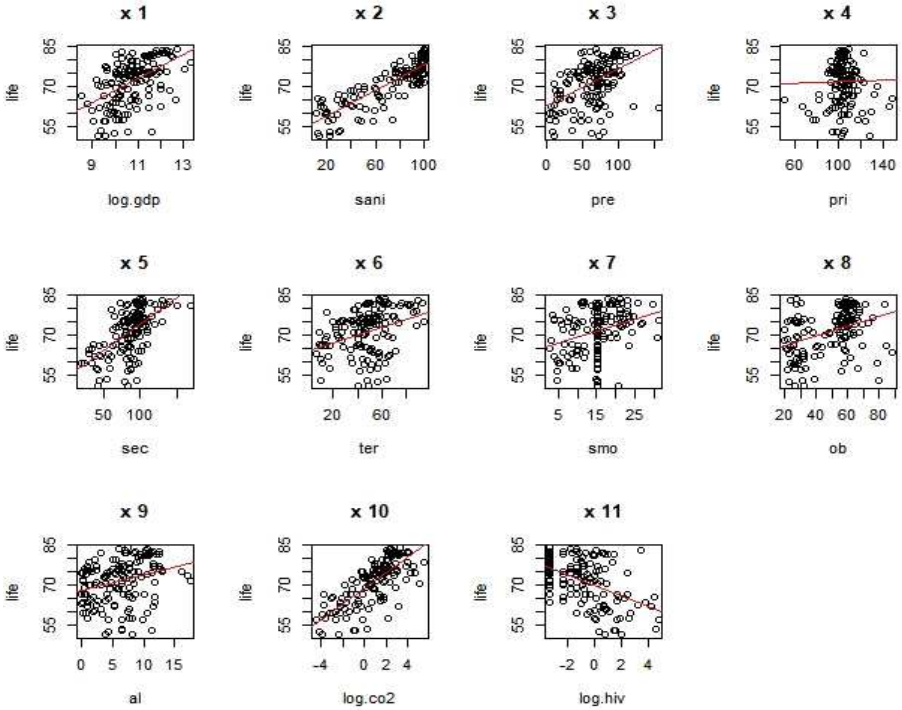
influential observation을 제거하기 전과 제거한 후 residual의 경향을 살펴보면, 제거한 후의 경향선의 기울기가 확실히 줄어든 것을 볼 수 있다. 이는 분산이 작아져서 거의 일정해졌다고 할 수 있다. 또한 QQ plot에서도 정규성 가정을 만족함을 알 수 있다.

4) Log transformation

위에서 여러 차례 언급했듯, heteroscedasticity 문제가 있으며 gdp 값이 너무 커서 rounding 했을 경우 coefficient가 0.0000으로 되는 문제도 발생한다. 아래는 각 변수별 산점도 이다.



gdp 변수와 co2, 그리고 hiv 변수의 경향성이 log를 취할 경우 선형성을 더 볼 수 있을 듯싶어, gdp 변수의 log 밑은 10, co2와 hiv 변수의 log 밑은 2로 하여 log transformation을 시행했다.



이 산점도를 살펴보면, log를 취해준 변수인 gdp와 co2, 그리고 hiv 중에서 선형 연관성이 이전보다 훨씬 강해졌음을 확인할 수 있다.

log transformation을 해주면, gdp는 0.0000에서 1.3420으로 gdp의 경향성을 확실하게 볼 수 있게 되었다. 마찬가지로 log.co2와 log.hiv 역시 경향성을 파악하기 더 쉬워졌다.

```
> round(obj.log$coefficients, 4)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.6191     5.3015  6.3414   0.0000
log.gdp       1.3420     0.3978  3.3735   0.0010
sani.1        0.1532     0.0195  7.8397   0.0000
pre.1         0.0317     0.0123  2.5677   0.0114
pri.1         0.0628     0.0279  2.2492   0.0262
sec.1         0.0167     0.0175  0.9571   0.3402
ter.1         0.0274     0.0177  1.5511   0.1233
smo.1        -0.1101     0.0572 -1.9233   0.0566
ob.1          0.0241     0.0206  1.1696   0.2443
al.1          0.1034     0.0882  1.1725   0.2431
log.co2       0.3554     0.2608  1.3629   0.1752
log.hiv      -0.6991     0.1737 -4.0250   0.0001
```

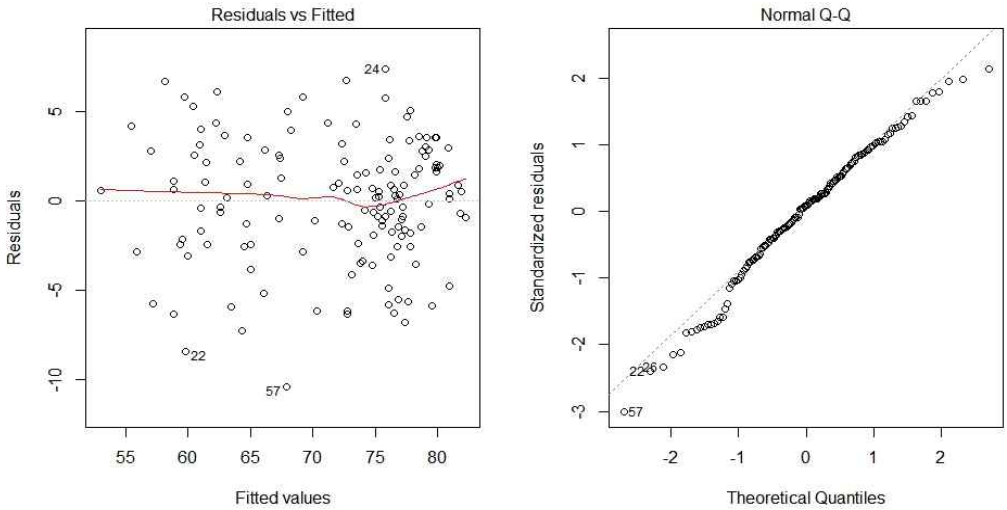
gdp와 co2, hiv에 log를 취해준 모델의 coefficient estimation은

$\widehat{\beta_0}= 33.6191, \widehat{\beta_1}= 1.3420, \widehat{\beta_2} = 0.1532, \widehat{\beta_3} = 0.0317, \widehat{\beta_4}= 0.0628, \widehat{\beta_5} = 0.0167, \widehat{\beta_6} = 0.0274,$
$\widehat{\beta_7}=-0.1101, \widehat{\beta_8}= 0.0241, \widehat{\beta_9}= 0.1034, \widehat{\beta_{10}}= 0.3554, \widehat{\beta_{11}}=-0.6991$

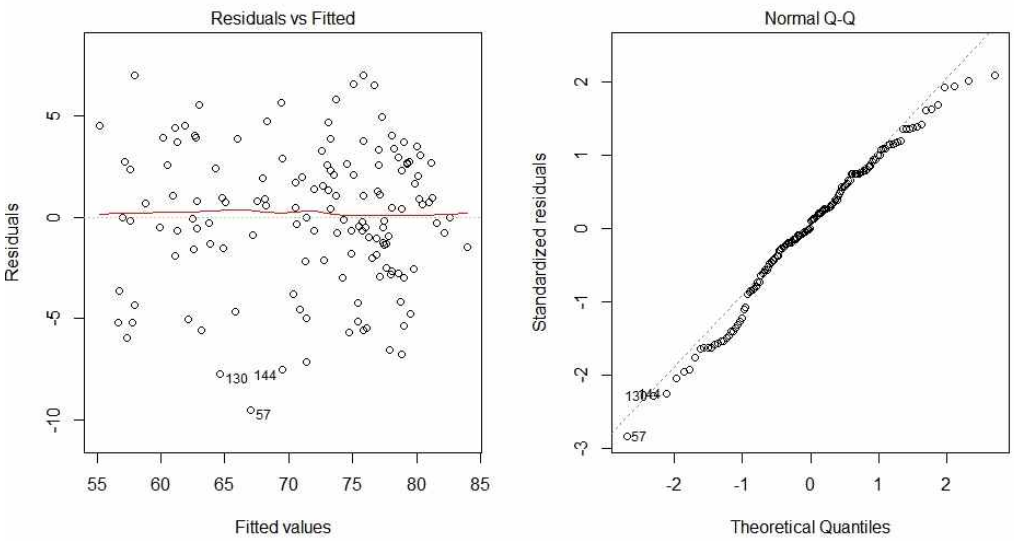
으로 추정할 수 있다.

다음은 log를 취해준 모델이다.

$$\widehat{life}_i = 33.6191 + 1.3420 \times \log.gdp_i + 0.1532 sani_i + 0.0317\, pre_i + 0.0628\, pri_i + 0.0167 \mathrm{sec}_i + 0.0385\, ter_i - 0.1101\, smo_i + 0.0241\, ob_i + 0.1034\, al_i + 0.3554 \log.co2_i - 0.6991 \log.hiv_i$$



〈log 변환 후, influential observation이 제거되지 않은 assumption test〉



〈log 변화 후, influential observation이 제거한 assumption test〉

log transformation을 한 이후의 heteroscedaticity 확인과 normality assumption을 확인한 결과이다. 일반 model의 residual plot을 보았을 때에는 heteroscedaticity가 다소 있었는데, log transformation을 취한 이후에는 그 정도가 약해진 것을 확인할 수 있다.

위는 influential observation을 제거하기 전의 residual plot과 QQ plot이고, 아래는 influential point obesrvation을 제거한 이후의 residual plot과 QQ plot 결과이다. Influential observation을 제거하기 전과 제거한 후 residual의 경향을 살펴보면, 제거한 후의 경향선의 기울기가 확실히 줄어든 것을 볼 수 있다. 이는 분산이 작아져서 거의 일정해졌다고 할 수 있다. 또한 QQ plot에서도 정규성 가정을 만족함을 알 수 있다.

다음은 log transformation model을 summary 함수를 통해 더 자세히 살펴본 결과이다.

```
> obg.log

Call:
lm(formula = life.1 ~ log.gdp + sani.1 + pre.1 + pri.1 + sec.1 +
    ter.1 + smo.1 + ob.1 + al.1 + log.co2 + log.hiv)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5180 -1.9624  0.1864  2.5660  6.9938

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.61911    5.30155   6.341 3.34e-09 ***
log.gdp      1.34195    0.39779   3.374 0.000974 ***
sani.1       0.15317    0.01954   7.840 1.34e-12 ***
pre.1        0.03168    0.01234   2.568 0.011353 *
pri.1        0.06283    0.02794   2.249 0.026154 *
sec.1        0.01671    0.01745   0.957 0.340250
ter.1        0.02740    0.01767   1.551 0.123261
smo.1       -0.11007    0.05723  -1.923 0.056599 .
ob.1         0.02408    0.02059   1.170 0.244260
al.1         0.10341    0.08820   1.173 0.243102
log.co2      0.35540    0.26077   1.363 0.175238
log.hiv     -0.69910    0.17369  -4.025 9.55e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.555 on 132 degrees of freedom
Multiple R-squared:  0.8243,    Adjusted R-squared:  0.8096
F-statistic: 56.29 on 11 and 132 DF,  p-value: < 2.2e-16
```

위에서 log transformation의 beta coefficients를 확인했던 바와 같이, 각 variable에 대한 추정치와 동시에 standard error, 그리고 각 coefficients에 대한 t-test를 수행한 결과를 확인해 볼 수 있다. $H_0 : \beta_j = 0 \ \ vs \ \ H_a : \beta_j \neq 0 \ \ (j = 1,2,\cdots,11)$하에서 가설검정을 하면 각 variable의 T-value를 구할 수 있다. 유의수준 0.05 수준에서 p-value를 살펴본 결과, log.gdp sanitation과 pre-primary, primary, 그리고 log hiv 모두 귀무가설을 기각 가능하다.

또한 모델 전체의 적합성을 보는 F-statistic이 56.29로 transformation 이전의 모델보다 커졌으며 역시 p-value가 0에 근사하므로 이 모델이 적합하다고 볼 수 있다. 게다가 Adjusted R^2가 0.8096으로() 설명력이 역시 높다.

따라서, log transformation model을 전체 data analysis에 사용할 수 있다고 판단한다.

5) Variable selection(Reduced model)

앞에서 도출한 다중회귀직선모형에서 모델의 수행 능력을 개선할 수 있는 최적의 변수들을 선택하도록 한다.

```
> obj$which
  (Intercept)   gdp sani   pre   pri   sec   ter smo.Daily.cigarettes    ob    al   co2   hiv
1        TRUE FALSE TRUE FALSE FALSE FALSE FALSE                FALSE FALSE FALSE FALSE FALSE
2        TRUE FALSE TRUE FALSE FALSE FALSE FALSE                FALSE FALSE FALSE FALSE  TRUE
3        TRUE FALSE TRUE  TRUE FALSE FALSE FALSE                FALSE FALSE FALSE FALSE  TRUE
4        TRUE  TRUE TRUE  TRUE FALSE FALSE FALSE                FALSE FALSE FALSE FALSE  TRUE
5        TRUE  TRUE TRUE  TRUE FALSE FALSE  TRUE                FALSE FALSE FALSE FALSE  TRUE
6        TRUE  TRUE TRUE  TRUE FALSE FALSE  TRUE                FALSE  TRUE FALSE FALSE  TRUE
7        TRUE  TRUE TRUE  TRUE  TRUE FALSE  TRUE                FALSE  TRUE FALSE FALSE  TRUE
8        TRUE  TRUE TRUE  TRUE  TRUE FALSE  TRUE                 TRUE  TRUE FALSE FALSE  TRUE
> obj$rsq
[1] 0.7333487 0.7591091 0.7866129 0.7958574 0.8020702 0.8075237 0.8130287 0.8171679
> which.max(obj$adjr2)
[1] 8
> which.min(obj$cp)
[1] 8
> which.min(obj$bic)
[1] 4
```

여기서, adjusted R^2 가 가장 높은 모델은 8번째로, sec, al, co2 변수가 빠지고, 나머지 8개의 변수로 구성된 다중회귀직선모형이 설명력이 높다.

다른 방법으로, Mallow's C_p 와 Bayesian/Schwarz Information Criterion(BIC/SIC)을 사용해보았다.

Mallow's C_p 는 $C_p = \frac{SSE_d}{MSE_d} + 2(d+1) - n$ 로 그 값이 작을수록 모델의 수행능력이 더 좋다고 볼 수 있다. Bayesian/Schwarz Information Criterion(BIC/SIC)도 마찬가지로 $n\log(SSE_d/n) + \log n \cdot d$ 값이 작을수록 모델의 수행능력이 더 좋다.

Mallow's C_p를 사용해보면 adjusted R^2 를 비교했을 때와 마찬가지로, 8번째 모델이 적합하고, BIC를 사용해보면, 4번째 모델, pri, sec, ter, smo, ob, al, co2가 빠지고, 나머지 4개의 변수로 구성된 다중회귀직선이 설명력이 높다.

먼저, BIC 방법을 사용해 도출한 reduced model은 다음과 같다.
$life_i = \beta_0 + \beta_1 gdp_i + \beta_2 sani_i + \beta_3 pre_i + \beta_{11} hiv_i$ 과 log transformation 전 기존의 full model과 이 reduced model을 비교해보았다.

```
> bic.1 <- lm(life.1~gdp.1+sani.1+pre.1+hiv.1)
> summary(bic.1)

Call:
lm(formula = life.1 ~ gdp.1 + sani.1 + pre.1 + hiv.1)

Residuals:
     Min       1Q   Median       3Q      Max
-12.3037  -2.3820   0.4512   2.7133   8.3653

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.387e+01  9.864e-01  54.613  < 2e-16 ***
gdp.1        7.043e-13  2.807e-13   2.509   0.0133 *
sani.1       2.046e-01  1.238e-02  16.520  < 2e-16 ***
pre.1        4.870e-02  1.175e-02   4.146 5.84e-05 ***
hiv.1       -3.416e-01  7.781e-02  -4.390 2.23e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.734 on 139 degrees of freedom
Multiple R-squared:  0.7959,    Adjusted R-squared:   0.79
F-statistic: 135.5 on 4 and 139 DF,  p-value: < 2.2e-16
```

여기서 살펴보면, BIC 방법을 사용하여 선택된 reduced model의 p-value가 0에 근사하므로, 적합하다고 판단할 수 있으며, 각 변수에 대한 t-test 결과 또한, 모두 유의하게 나왔다.

이 결과를 바탕으로, 더 적절한 model 선택을 위하여 full model과 reduced model을 비교해보았다. $H_0 : \beta_j = 0 \ \ vs \ \ H_a : \ \text{not} \ \ H_0 \qquad j = 4, 5, 6, 7, 8, 9, 10$ 가정에서 global test를 수행한 결과는 아래와 같다.

```
> anova(bic.1, reg.1)
Analysis of Variance Table

Model 1: life.1 ~ gdp.1 + sani.1 + pre.1 + hiv.1
Model 2: life.1 ~ gdp.1 + sani.1 + pre.1 + pri.1 + sec.1 + ter.1 + smo.1 +
    ob.1 + al.1 + co2.1 + hiv.1
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    139 1938.4
2    132 1705.8  7    232.59 2.5711 0.01623 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value를 보면 유의수준 0.05에서 기각이 가능하며, 따라서, full model이 BIC 방법을 통해 도출한 reduced model보다 더 좋은 모델이라 평가할 수 있다.

반면, Mallow's C_p 방법을 사용해 도출한 reduced model은 다음과 같다.
$life_i = \beta_0 + \beta_1 gdp_i + \beta_2 sani_i + \beta_3 pre_i + \beta_4 pri_i + \beta_6 ter_i + \beta_7 smo_i + \beta_8 ob_i + \beta_{11} hiv_i$ 을 선택해서 log transformation 전 기존의 full model과 이 reduced model을 비교해보았다.

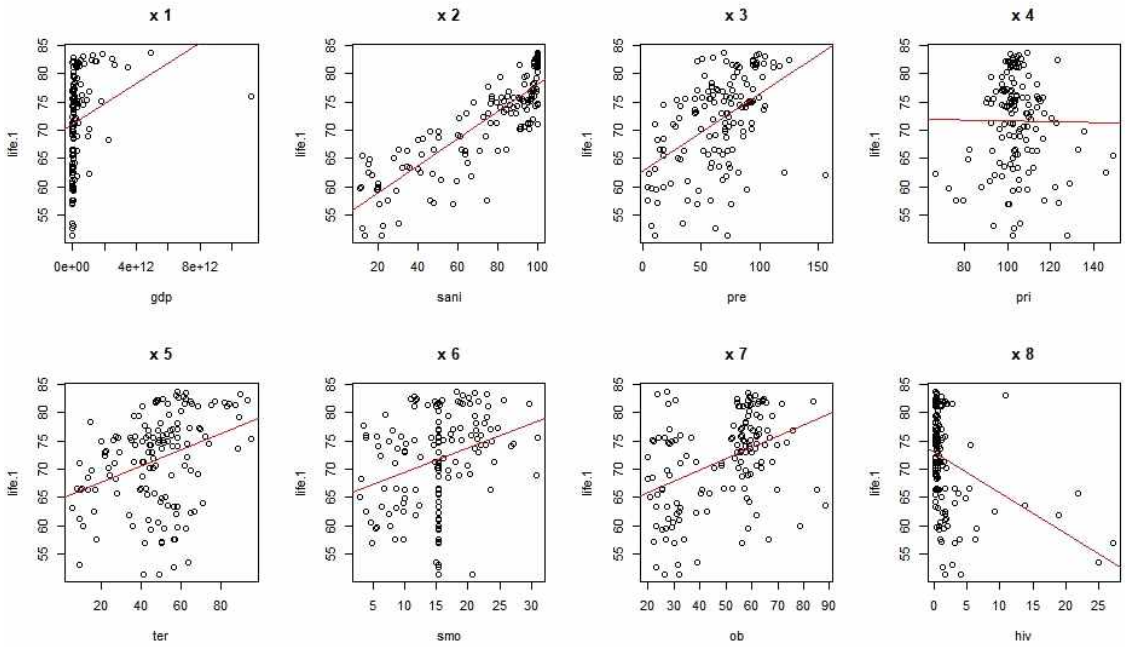
```
> summary(reg.cp)

Call:
lm(formula = life.1 ~ gdp.1 + sani.1 + pre.1 + pri.1 + ter.1 +
    smo.1 + ob.1 + hiv.1)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1796  -1.9384   0.2931   2.4806   7.4366

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.635e+01  3.304e+00  14.027  < 2e-16 ***
gdp.1        8.071e-13  2.758e-13   2.926  0.00403 **
sani.1       1.969e-01  1.368e-02  14.385  < 2e-16 ***
pre.1        3.850e-02  1.173e-02   3.281  0.00132 **
pri.1        5.581e-02  2.743e-02   2.035  0.04384 *
ter.1        4.663e-02  1.693e-02   2.754  0.00670 **
smo.1       -9.789e-02  5.599e-02  -1.748  0.08270 .
ob.1         4.537e-02  2.076e-02   2.185  0.03058 *
hiv.1       -3.750e-01  7.566e-02  -4.956 2.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.586 on 135 degrees of freedom
Multiple R-squared:  0.8172,    Adjusted R-squared:  0.8063
F-statistic: 75.42 on 8 and 135 DF,  p-value: < 2.2e-16
```

여기서 살펴보면, Mallow's C_p 방법을 사용하여 선택된 reduced model의 p-value가 0에 근사하므로, 적합하다고 판단할 수 있으며, 각 변수에 대한 t-test 결과도 한 변수 smo만 제외하면 모두 유의하게 나왔다.

이 결과를 바탕으로, 더 적절한 model 선택을 위하여 full model과 reduced model을 비교해보았다. $H_0 : \beta_j = 0 \ \ vs \ \ H_a$: not H_0 $\quad j = 5, 9, 10$ 가정에서 global test를 수행한 결과는 아래와 같다.

```
> reg.cp <- lm(life.1~gdp.1+sani.1+pre.1+pri.1+ter.1+smo.1+ob.1+hiv.1)
> anova(reg.cp, reg.1)
Analysis of Variance Table

Model 1: life.1 ~ gdp.1 + sani.1 + pre.1 + pri.1 + ter.1 + smo.1 + ob.1 +
    hiv.1
Model 2: life.1 ~ gdp.1 + sani.1 + pre.1 + pri.1 + sec.1 + ter.1 + smo.1 +
    ob.1 + al.1 + co2.1 + hiv.1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    135 1736.1
2    132 1705.8  3    30.234 0.7798 0.5072
```

p-value를 보면 유의수준 0.05에서 기각을 할 수 없다, 따라서, Mallow's C_p 방법을 통해 도출한 reduced model이 full model보다 더 좋은 모델이라 평가할 수 있다.



Mallow's C_p 방법을 통해 도출한 reduced model에서 각 explanatory variable과 response variable 간의 scatter plot을 그려본 결과이다.

다만, 계속 언급했듯이, 일반 모델에 있어서 gdp 변수의 값이 너무 커서 rounding 할 경우 coefficient가 0이 되는 문제와 residual plot에서 heteroscedaticity가 보이는 문제가 있었다. 그래서 log transformation을 하고 난 뒤의 variable selection을 다시 한 번 수행해 보도록 하였다.

```
> obj$which
  (Intercept) log.gdp sani.1 pre.1 pri.1 sec.1 ter.1 smo.1  ob.1  al.1 log.co2 log.hiv
1        TRUE   FALSE   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE   FALSE
2        TRUE    TRUE   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE   FALSE
3        TRUE    TRUE   TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE   FALSE
4        TRUE    TRUE   TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE    TRUE
5        TRUE    TRUE   TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE   FALSE    TRUE
6        TRUE    TRUE   TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE    TRUE    TRUE
7        TRUE    TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE    TRUE    TRUE
8        TRUE    TRUE   TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE    TRUE    TRUE
> obj$rsq
[1] 0.7333487 0.7665239 0.7850151 0.8021240 0.8064079 0.8110217 0.8150416 0.8189571
> which.max(obj$adjr2)
[1] 8
> which.min(obj$cp)
[1] 8
> which.min(obj$bic)
[1] 4
```

이는 log transformation 모델의 variable selection과정이다. 위와 동일한 방법으로, Adjusted R^2, mallow C_p 그리고 BIC method를 사용해 variable을 선택한 결과이다. Adjusted R^2에 따른 모델은 위의 1~8개의 모델 중 log.gdp, sanitation, pre-primary, primary, tertiary, smoking, log.co2, 그리고 log.hiv가 선택되었다. 또한 mallow C_p criterion에 의한 결과도 위와 동일하다. 다만 BIC를 사용했을 경우 4번째 log.gdp, sanitation, pre-primary, 그리고 log.hiv를 predictor로 놓는 모델이 선택되었다. 일단 variable이 가장 작은 BIC model을 가지고 모델의 적합성을 판단해 보기로 한다.

$\widehat{life} = \widehat{\beta_0} + \widehat{\beta_1}\log.gdp_i + \widehat{\beta_2}sani_i + \widehat{\beta_3}pre_i + \widehat{\beta_{11}}\log.hiv_i$

```
> bic.log <- lm(life.1~log.gdp+sani.1+pre.1+log.hiv)
> summary(bic.log)

Call:
lm(formula = life.1 ~ log.gdp + sani.1 + pre.1 + log.hiv)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9156  -2.1390   0.2425   2.5713   8.0169

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.10794    3.83801   9.929  < 2e-16 ***
log.gdp      1.56202    0.38906   4.015 9.69e-05 ***
sani.1       0.18150    0.01375  13.199  < 2e-16 ***
pre.1        0.04677    0.01164   4.017 9.60e-05 ***
log.hiv     -0.58446    0.16859  -3.467 0.000701 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.677 on 139 degrees of freedom
Multiple R-squared:  0.8021,    Adjusted R-squared:  0.7964
F-statistic: 140.9 on 4 and 139 DF,  p-value: < 2.2e-16
```

BIC log transformed Reduced model에 대한 summary 과정이다. 모든 variable에 대한 t-test 결과는 0.05 수준에서 모두 귀무가설을 기각 가능하다. 따라서 이 모델에서 사용된 모든 변수는 사용할 수 있다고 판단한다. 또한, 전체 모델의 적합성을 판단하는 F-test 결과를 보면, F statistic 값이 140.9이고 P-value가 0에 근사하므로 모델 자체는 적합성이 있다고 판단한다. 그렇다면 과연 이 log transformed Reduced Model과 log transformed Full Model 중 무슨 모델이 더 적합한지에 대해 가설검정을 진행해 보도록 한다. 그 결과가 아래와 같다.

```
> anova(bic.log, reg.log)
Analysis of Variance Table

Model 1: life.1 ~ log.gdp + sani.1 + pre.1 + log.hiv
Model 2: life.1 ~ log.gdp + sani.1 + pre.1 + pri.1 + sec.1 + ter.1 + smo.1 +
    ob.1 + al.1 + log.co2 + log.hiv
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    139 1878.9
2    132 1668.5  7    210.39 2.3778 0.02541 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \beta_j = 0 \;\; vs \;\; H_a :$ not $H_0 \qquad j = 4, 5, 6, 7, 8, 9, 10$

귀무가설은 log transformed Reduced Model이며 대립가설은 log transformed Full Model이다. 이 가설검정을 수행했을 때, BIC log transformed Reduced Model이 P-value가 0.02541로 작게 나오므로 유의수준 0.05 수준에서 귀무가설을 기각할 수 있다. 따라서 Reduced Model을 사용하는 것보다 Full Model을 사용하는 것이 낫다는 결과가 도출된다. 따라서 이 BIC log transformed Reduced Model을 사용하는 것 대신 C_p criterion을 사용한 model을 사용하도록 한다.

$\widehat{life} = \widehat{\beta_0} + \widehat{\beta_1} gdp_i + \widehat{\beta_2} sani_i + \widehat{\beta_3} pre_i + \widehat{\beta_4} pri_i + \widehat{\beta_6} ter_i + \widehat{\beta_7} smo_i + \widehat{\beta_{10}} co2_i + \widehat{\beta_{11}} hiv_i$

```
> summary(reg.logcp)

Call:
lm(formula = life.1 ~ log.gdp + sani.1 + pre.1 + pri.1 + ter.1 +
    smo.1 + log.co2 + log.hiv)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2827  -2.1695   0.3725   2.3685   8.2023

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.36378    5.10420   6.928 1.58e-10 ***
log.gdp      1.31049    0.39642   3.306  0.00121 **
sani.1       0.16301    0.01888   8.634 1.46e-14 ***
pre.1        0.03836    0.01175   3.265  0.00139 **
pri.1        0.06474    0.02761   2.344  0.02052 *
ter.1        0.03305    0.01699   1.945  0.05380 .
smo.1       -0.09528    0.05576  -1.709  0.08980 .
log.co2      0.46852    0.25394   1.845  0.06723 .
log.hiv     -0.62038    0.16487  -3.763  0.00025 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*다시

```
Residual standard error: 3.568 on 135 degrees of freedom
Multiple R-squared:  0.819,     Adjusted R-squared:  0.8082
F-statistic: 76.33 on 8 and 135 DF,  p-value: < 2.2e-16
```

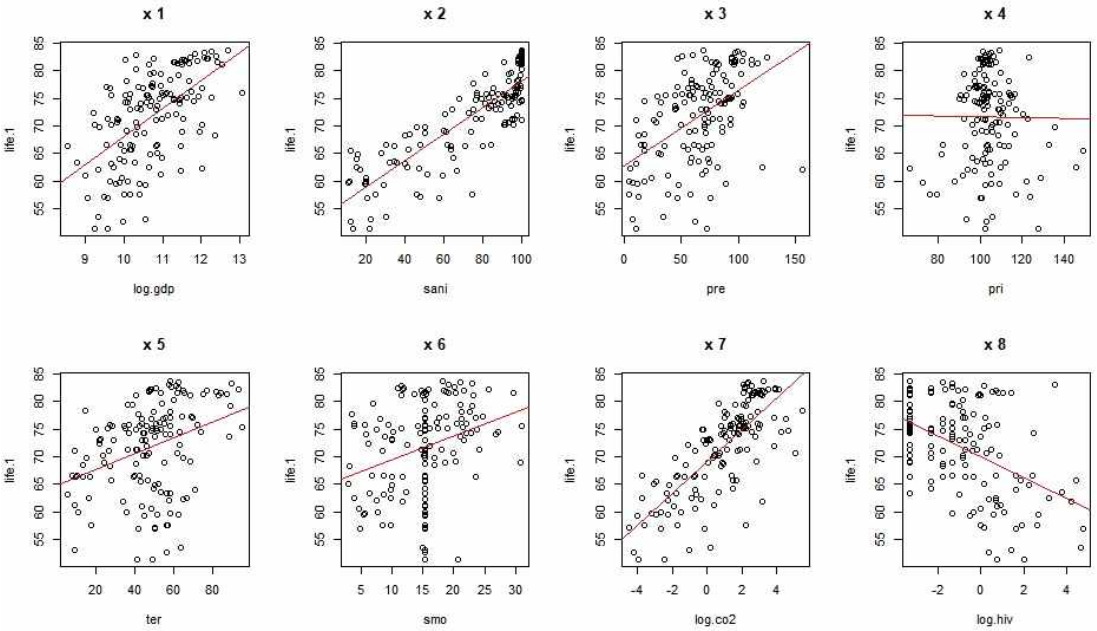
Mallow log transformed C_p Reduced model에 대한 summary 과정이다. 대부분의 variable에 대한 t-test 결과는 0.05 수준에서 모두 귀무가설을 기각 가능하다. 또한, 전체 모델의 적합성을 판단하는 F-test 결과를 보면, F statistic 값이 76.33이고 P-value가 0에 근사하므로 모델 자체는 적합성이 있다고 판단한다. 그렇다면 과연 이 log transformed Reduced Model과 log transformed Full Model 중 무슨 모델이 더 적합한지에 대해 가설검정을 진행해 보도록 한다. 그 결과가 아래와 같다.

```
> reg.logcp<- lm(life.1~log.gdp+sani.1+pre.1+pri.1+ter.1+smo.1+log.co2+log.hiv)
> anova(reg.logcp, reg.log)
Analysis of Variance Table

Model 1: life.1 ~ log.gdp + sani.1 + pre.1 + pri.1 + ter.1 + smo.1 + log.co2 +
    log.hiv
Model 2: life.1 ~ log.gdp + sani.1 + pre.1 + pri.1 + sec.1 + ter.1 + smo.1 +
    ob.1 + al.1 + log.co2 + log.hiv
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    135 1719.1
2    132 1668.5  3    50.554 1.3331 0.2664
```

$H_0 : \beta_j = 0 \quad vs \quad H_a :$ not $H_0 \qquad j = 5, 8, 9$

귀무가설은 log transformed Reduced Model이며 대립가설은 log transformed Full Model이다. 이 가설검정을 수행했을 때, BIC Reduced Model과는 다르게 P-value가 0.2664로 크게 나오므로 귀무가설을 기각할 수 없다. 따라서 log transformed Full Model을 사용하는 것보다 log transformed Reduced Model을 사용하는 것이 더 용이함을 알 수 있다. 그래서 mallow C_p log Reduced Model을 사용하기로 한다.

아래는 선택된 변수들에 대한 scatter plot이다.



log를 취한 후의 gdp, co2, hiv에서 상당히 선형성이 강해진 것을 확인할 수 있다.

6) Multiple collinearity

Mallow's C_p 방법을 통해 도출한 reduced model의 다중공선성을 vif() 함수를 사용해서 확인해보면 다음과 같다. 일반적인 연구에서는 10보다 작은 경우는, 다중공선성이 없다고 판단한다. 따라서 이 reduced model에서는 다중공선성이 발견되지 않는다.

```
> vif(reg.r)
  gdp.1 sani.1  pre.1  pri.1  ter.1  smo.1   ob.1  hiv.1
1.0823 1.7343 1.3246 1.0479 1.1998 1.2177 1.3395 1.1018
```

log transformation 후, Mallow's C_p 방법을 통해 도출한 reduced model의 다중공선성을 vif() 함수를 사용해서 확인해보면 다음과 같다. 일반적인 연구에서는 10보다 작은 경우는, 다중공선성이 없다고 판단한다. log transformation 전의 reduced model보다는 수치가 더 크긴 하지만, 이 reduced model도 다중공선성이 발견되지 않는다.

```
> vif(reg.logr)
log.gdp  sani.1   pre.1   pri.1   ter.1   smo.1 log.co2 log.hiv
 1.4608  3.3332  1.3417  1.0724  1.2201  1.2195  3.3959  1.2589
```

7) Model selection

이제 위에서 일반 모델과 log transformed 된 후의 모델 중에서 더 적합한 모델을 선택하기 위해서 AIC와 BIC를 사용해 보았다. 이 두 방법은 적합성(Goodness of fit)을 측정해주는 지표이다. AIC(Akaike Information Criterion)는 $AIC = nlog(SSE_d/n) + 2 \cdot d$ 로 표현한다. SSE_d는 reduced model의 SSE이며 d는 reduced model의 predictor을 의미한다. BIC(Bayesian/Schwarz Information Criterion)는 $BIC = nlog(SSE_d/n) + \log n \cdot d$ 로 표현된다.

이 두 방법을 수행한 결과가 작을수록 더 나은 performance를 의미한다.

```
> AIC(reg.r, reg.logr)
         df      AIC
reg.r    10 787.1532
reg.logr 10 785.7371
> BIC(reg.r, reg.logr)
         df      BIC
reg.r    10 816.8514
reg.logr 10 815.4353
```

transformation을 하기 전의 Reduced Model과 log transformed을 한 후의 Reduced Model을 비교해 본 결과, AIC와 BIC 모두 미세하게나마 log transformed 된 후의 Reduced Model이 더 나은 Performance를 보이는 것으로 나타났다. 따라서 log transformed Reduced Model을 사용하도록 결정하였다.

Ⅲ. 결론

본 연구의 목적은 다중회귀 분석과 관련된 실증분석을 통해 새로운 현상을 발견하기보다 다중회귀 분석을 통해서 적절한 모델을 선택하는 데에 있다.

1. 분석 결과

$$\widehat{life} = \widehat{\beta_0} + \widehat{\beta_1}\log_{10}gdp_i + \widehat{\beta_2}sani_i + \widehat{\beta_3}pre_i + \widehat{\beta_4}pri_i + \widehat{\beta_6}ter_i + \widehat{\beta_7}smo_i + \widehat{\beta_{10}}log_2co2_i + \widehat{\beta_{11}}log_2hiv_i$$

model에 따르면, life expectancy at birth에는 gdp, sanitation, pre-primary, primary, tertiary, smoking rate, CO2 emissions, and hiv rate 이 8개의 변수가 유의미한 영향을 미치는 것으로 볼 수 있다. secondary enrollment ratio와 obesity, 그리고 alcohol 변수는 life expectancy에 크게 관련이 없다고 판단한다.

```
> summary(reg.logcp)

Call:
lm(formula = life.1 ~ log.gdp + sani.1 + pre.1 + pri.1 + ter.1 +
    smo.1 + log.co2 + log.hiv)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2827  -2.1695   0.3725   2.3685   8.2023

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.36378    5.10420   6.928 1.58e-10 ***
log.gdp      1.31049    0.39642   3.306  0.00121 **
sani.1       0.16301    0.01888   8.634 1.46e-14 ***
pre.1        0.03836    0.01175   3.265  0.00139 **
pri.1        0.06474    0.02761   2.344  0.02052 *
ter.1        0.03305    0.01699   1.945  0.05380 .
smo.1       -0.09528    0.05576  -1.709  0.08980 .
log.co2      0.46852    0.25394   1.845  0.06723 .
log.hiv     -0.62038    0.16487  -3.763  0.00025 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.568 on 135 degrees of freedom
Multiple R-squared:  0.819,     Adjusted R-squared:  0.8082
F-statistic: 76.33 on 8 and 135 DF,  p-value: < 2.2e-16
```

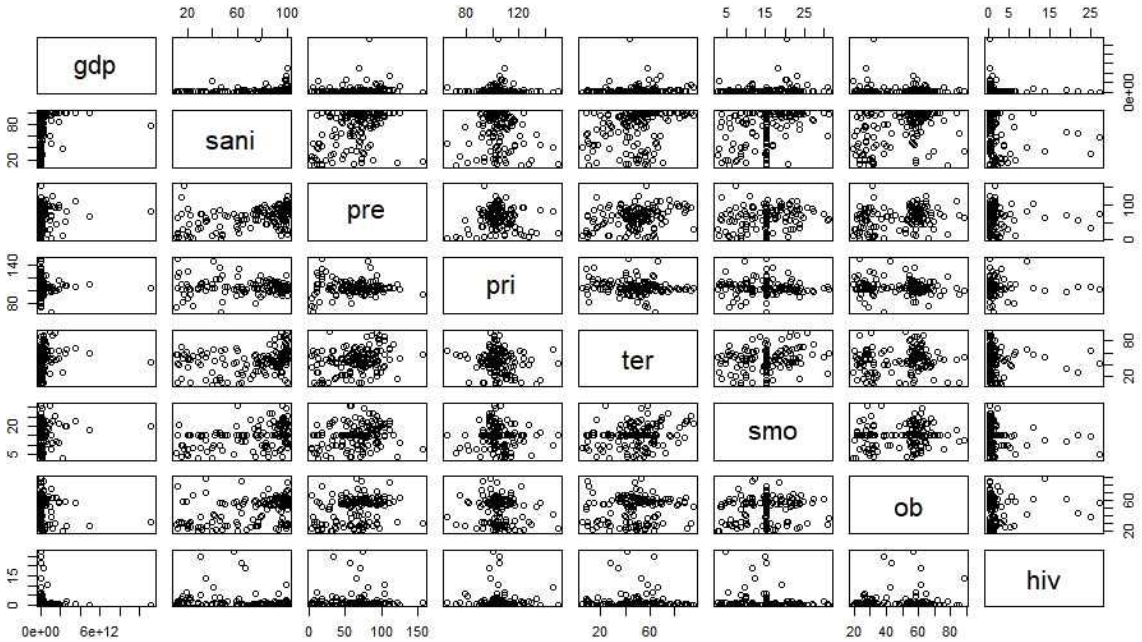
따라서, 1인당 실질 GDP나 향상된 위생시설(sanitation) 확충, 교육수준 평가 지표 중 유치원 등록률, 고등학교 등록률 등은 기대 수명과 양의 상관관계를 갖는 것으로 보인다. 흡연율과 HIV 출현율은 coefficient가 음수로, 기대 수명과 음의 상관관계를 갖는 것으로 보인다.

2. 한계점 및 의문점

우선, 자료 수집에 있어서, 출처가 다른 곳에서 데이터를 수집하다 보니 변수마다 수집된 자료의 년도가 달랐다. 그래서 각 변수별로 시점을 하나로 통일할 수 없었다는 한계가 있다. 또한, 분석 순서에 있어서 transformation 된 이후의 model에서 influential point를 제거하는 대신 기존 모델에서 influential point를 제거한 후에 log transformation을 취해주었는데 그 결과가 어떻게 변화하는지 확인하지 못하였다.

Assumption test 중 heteroscedaticity를 점검하는 과정에서 본래는 WLO를 사용해야 했으나 그 대신 transformation을 해 주었다.

Variable selection 과정에서 BIC, mallow C_p 방법을 사용하였는데 무슨 이유로 Full Model과 비교하는 과정에서 결과가 다른지 궁금하다.



위의 각 설명변수 간의 산점도를 보면 obesity 변수에 cluster가 두 개 있음을 알 수 있다. 설명변수에 cluster가 있을 때의 분석 방법에 대하여 탐구해 볼 필요가 있다.

뽑은 최적의 모델인 log transformed Reduced Model에서 smoking 변수의 coefficient estimator은 음수이나 기대 수명과 1대 1 비교한 scatter plot에서 확인한 single linear model은 양의 기울기를 가졌다. 그 이유에 대해 탐구해 볼 필요가 있다.

마지막으로, 개선할 점은 시간상 국가별로 기대 수명과 사회적 요인들 간의 관계를 살펴보지 못했다. 차후에 이런 부분을 더 추가해 볼 예정이다.

참고자료

1. 김명중, 박범조. R을 이용한 분위수회귀 분석 : 경제외적 요인이 기대수명에 미치는 영향, DKU 미래산업연구소_ 단국대학교 산업연구 37권 2호, 2013, p33-68.
2. 최용옥, 급속한 기대수명 증가의 함의(Longevity Risk in Korea), KDI FOCUS, NO.69, (Korea Development Institute), 2016, p3
3. https://data.worldbank.org/indicator/SP.DYN.LE00.IN
4. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
5. https://data.worldbank.org/indicator/SH.STA.ACSN?name_desc=false
6. https://data.worldbank.org/indicator/SE.PRE.ENRR
7. https://data.worldbank.org/indicator/SE.PRM.ENRR
8. https://data.worldbank.org/indicator/SE.SEC.ENRR
9. https://data.worldbank.org/indicator/SE.TER.ENRR
10. http://apps.who.int/gho/data/view.main.TOB30011
11. http://apps.who.int/gho/data/view.main.CTRY2430A
12. https://data.worldbank.org/indicator/SH.ALC.PCAP.LI
13. https://data.worldbank.org/indicator/EN.ATM.CO2E.PC
14. https://data.worldbank.org/indicator/EN.ATM.CO2E.PC
15. https://data.humdata.org/dataset/prevalence_of_hiv_total_of_population_aged_15-49/resource/c5f56338-471b-4aaf-b5b9-b1f7db160bc1
16. http://www.datamarket.kr/xe/board_BoGi29/9880
17. http://www.saedsayad.com/k_nearest_neighbors_reg.htm