과목명:	기계학습 및 실습
과제번호:	기말 프로젝트
담당교수:	박창이 교수님
제출일:	**2020년 6월26일(금)**
학과:	**통계학과**
학번:	**2018580004**
이름:	**김 다 희**

서울시립대학교
UNIVERSITY OF SEOUL

- 목차 -

1. 서론

1. [1]데이터에 대한 설명

선택한 데이터는 세계보건기구(WHO)에서 제공한 데이터셋이다. 총 193개국의 2000~2015년 인구의 기대수명과 건강에 영향을 미치는 다양한 요인 중 대표적으로 중요한 요인에 대한 정보를 포함하고 있다. 총 2,938개의 관측과 22개의 변수를 가지고 있으며 이때 단순한 건강 관련 요인뿐만 아니라 경제적, 사회적 요인 등 다양한 관점의 정보를 가지고 있다. 데이터 초기 검사 과정에서 일부 결측값이 포함되었는데 다행히 큰 오류는 발견되지 않았다고 한다. 따라서 결측값만 잘 처리해서 기대수명에 미치는 건강 요인 분석해보고자 한다.

2. 분석 목표

데이터 분석 목표는 기대수명에 영향을 미치는 요인을 찾고 어떤 변수가 더 영향력 있는지 알아보는 것이다. 기대수명에 영향을 미치는 요인을 제대로 파악하면 나라에서 인구의 기대수명을 증가시키기 위한 계획을 효율적으로 짤 수 있어 개인뿐만 아니라 국가의 경제적으로도 좋은 영향을 미칠 수 있다. 따라서 반응변수를 기대수명으로 하는 선형모형을 바탕으로 다양한 방법을 통해 변수를 선택하고 모형을 적합하여 기대수명과 관련된 설명변수를 찾아보겠다.

---

1) 데이터 출처: https://www.kaggle.com/kumarajarshi/life-expectancy-who (캐글)

2. 본론

1. 데이터에 대한 기초 탐색

먼저 데이터 분석을 위해 위에서 설명한 기대수명과 건강 관련 요인에 대한 데이터 셋을 R에 불러오고 본격적인 데이터 분석에 앞서 데이터에 대해서 살펴보자.

(1) 데이터 불러오기

```
> who = read.csv("LifeExpectancy.csv", header=T, na.strings = "?")
```

(2) head(who)

```
> head(who) # 앞에서부터 6행까지의 데이터 확인
      Country Year     Status Life.expectancy Adult.Mortality infant.deaths Alcohol percentage.expenditure
1 Afghanistan 2015 Developing            65.0             263            62    0.01              71.279624
2 Afghanistan 2014 Developing            59.9             271            64    0.01              73.523582
3 Afghanistan 2013 Developing            59.9             268            66    0.01              73.219243
4 Afghanistan 2012 Developing            59.5             272            69    0.01              78.184215
5 Afghanistan 2011 Developing            59.2             275            71    0.01               7.097109
6 Afghanistan 2010 Developing            58.8             279            74    0.01              79.679367
  Hepatitis.B Measles  BMI under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
1          65    1154 19.1                83     6               8.16         65      0.1 584.25921   33736494
2          62     492 18.6                86    58               8.18         62      0.1 612.69651     327582
3          64     430 18.1                89    62               8.13         64      0.1 631.74498   31731688
4          67    2787 17.6                93    67               8.52         67      0.1 669.95900    3696958
5          68    3013 17.2                97    68               7.87         68      0.1  63.53723    2978599
6          66    1989 16.7               102    66               9.20         66      0.1 553.32894    2883167
  thinness..1.19.years thinness.5.9.years Income.composition.of.resources Schooling
1                 17.2                17.3                          0.479      10.1
2                 17.5                17.5                          0.476      10.0
3                 17.7                17.7                          0.470       9.9
4                 17.9                18.0                          0.463       9.8
5                 18.2                18.2                          0.454       9.5
6                 18.4                18.4                          0.448       9.2
```

⇒ **앞의 6개의 관측을 통해 데이터의 생김새를 대략 살펴볼 수 있다.**

(3) dim(who)

```
> dim(who) # 2938개의 관측값, 22개의 변수
[1] 2938   22
```

⇒ **2938개의 관측값(행), 22개의 변수(열)를 가지고 있다.**

(4) names(who)

```
> names(who) # 22개의 변수명 확인
 [1] "Country"                  "Year"                     "Status"
 [4] "Life.expectancy"          "Adult.Mortality"          "infant.deaths"
 [7] "Alcohol"                  "percentage.expenditure"   "Hepatitis.B"
[10] "Measles"                  "BMI"                      "under.five.deaths"
[13] "Polio"                    "Total.expenditure"        "Diphtheria"
[16] "HIV.AIDS"                 "GDP"                      "Population"
[19] "thinness..1.19.years"     "thinness.5.9.years"       "Income.composition.of.resources"
[22] "Schooling"
```

⇒ **기대수명과 이에 관련된 여러 요인에 대한 총 22개의 변수 명을 확인할 수 있다. 단순히 HIV 바이러스, 음주 등과 같은 건강과 밀접한 요인뿐만 아니라 GDP, 학업 수준과 같은 사회적 요인도 포함하고 있다.**

(4) str(who)

```
> str(who) # 데이터의 구조 확인; Country, Status만 chr
'data.frame':   2938 obs. of  22 variables:
 $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
 $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
 $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
 $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
 $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
 $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

⇒ 데이터의 구조를 확인할 수 있다. 22개의 변수 중 Country, Status만 문자형 변수고 나머지는 모두 수치형 변수임을 알 수 있다.

(5) summary(who)

```
> summary(who) # 데이터의 요약 통계량 확인
   Country               Year         Status          Life.expectancy Adult.Mortality infant.deaths
 Length:2938        Min.   :2000   Length:2938        Min.   :36.30   Min.   :  1.0   Min.   :   0.0
 Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10   1st Qu.: 74.0   1st Qu.:   0.0
 Mode  :character   Median :2008   Mode  :character   Median :72.10   Median :144.0   Median :   3.0
                    Mean   :2008                      Mean   :69.22   Mean   :164.8   Mean   :  30.3
                    3rd Qu.:2012                      3rd Qu.:75.70   3rd Qu.:228.0   3rd Qu.:  22.0
                    Max.   :2015                      Max.   :89.00   Max.   :723.0   Max.   :1800.0
                                                      NA's   :10      NA's   :10
    Alcohol        percentage.expenditure  Hepatitis.B       Measles               BMI        under.five.deaths
 Min.   : 0.0100   Min.   :    0.000      Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
 1st Qu.: 0.8775   1st Qu.:    4.685      1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
 Median : 3.7550   Median :   64.913      Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
 Mean   : 4.6029   Mean   :  738.251      Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
 3rd Qu.: 7.7025   3rd Qu.:  441.534      3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
 Max.   :17.8700   Max.   :19479.912      Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
 NA's   :194                              NA's   :553                        NA's   :34
     Polio       Total.expenditure   Diphtheria       HIV.AIDS           GDP               Population
 Min.   : 3.00   Min.   : 0.370   Min.   : 2.00   Min.   : 0.100   Min.   :     1.68   Min.   :3.400e+01
 1st Qu.:78.00   1st Qu.: 4.260   1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   463.94   1st Qu.:1.958e+05
 Median :93.00   Median : 5.755   Median :93.00   Median : 0.100   Median :  1766.95   Median :1.387e+06
 Mean   :82.55   Mean   : 5.938   Mean   :82.32   Mean   : 1.742   Mean   :  7483.16   Mean   :1.275e+07
 3rd Qu.:97.00   3rd Qu.: 7.492   3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  5910.81   3rd Qu.:7.420e+06
 Max.   :99.00   Max.   :17.600   Max.   :99.00   Max.   :50.600   Max.   :119172.74   Max.   :1.294e+09
 NA's   :19      NA's   :226      NA's   :19                       NA's   :448         NA's   :652
 thinness..1.19.years thinness.5.9.years Income.composition.of.resources   Schooling
 Min.   : 0.10        Min.   : 0.10      Min.   :0.0000                   Min.   : 0.00
 1st Qu.: 1.60        1st Qu.: 1.50      1st Qu.:0.4930                   1st Qu.:10.10
 Median : 3.30        Median : 3.30      Median :0.6770                   Median :12.30
 Mean   : 4.84        Mean   : 4.87      Mean   :0.6276                   Mean   :11.99
 3rd Qu.: 7.20        3rd Qu.: 7.20      3rd Qu.:0.7790                   3rd Qu.:14.30
 Max.   :27.70        Max.   :28.60      Max.   :0.9480                   Max.   :20.70
 NA's   :34           NA's   :34         NA's   :167                      NA's   :163
```

⇒ 마지막으로 summary 함수를 통해 각 변수의 요약 통계량까지 확인해보았다. 눈에 띄는 점은 데이터 설명에서 언급했듯이 NA가 꽤 많이 포함된 데이터인 듯하다. 이어지는 데이터 전처리 단계에서 이런 결측값을 처리하도록 하겠다.

2. 데이터 전처리

(1) "Country" 변수 제거

```
> who = who[,-1] # 변수 제거
> names(who) # 확인
 [1] "Year"                       "Status"                         "Life.expectancy"
 [4] "Adult.Mortality"            "infant.deaths"                  "Alcohol"
 [7] "percentage.expenditure"     "Hepatitis.B"                    "Measles"
[10] "BMI"                        "under.five.deaths"              "Polio"
[13] "Total.expenditure"          "Diphtheria"                     "HIV.AIDS"
[16] "GDP"                        "Population"                     "thinness..1.19.years"
[19] "thinness.5.9.years"         "Income.composition.of.resources" "Schooling"
```

⇒ 먼저 문자형 변수는 회귀분석하기 까다로우며 기대수명과 국가의 영향을 직접적으로 분석하고자 하는 게 아니기 때문에 제외하였다. 또한 GDP가 국가의 역할을 어느 정도 대체할 수 있다. 따라서 Country 변수 제거해 총 1개의 예측변수(기대수명)와 20개의 설명변수를 가진 데이터가 되었다.

(2) "Status" 변수 변환

```
> attach(who)
> Status = as.factor(Status)
> class(Status)
[1] "factor"
```

⇒ Status는 국가의 개발 상태를 나타내는 변수로 Developing(개발도상국)과 Developed(선진국) 두 수준으로 나뉜다. 원활한 분석을 위해 factor형으로 바꿔주었다.

(3) 결측값 처리

```
> sum(is.na(who)) # 결측치 개수 확인
[1] 2563
> mean(is.na(who)) # 결측치가 꽤 많은 비중; 2938*22칸 중 2563개의 칸이 비어있음
[1] 0.04154105
> mean(!complete.cases(who)) # 적어도 한 개 이상의 결측값을 가지는 행; 거의 반 정도 결측치 포함
[1] 0.4387338
```

⇒ 결측값을 모두 제거하기엔 데이터 손실이 너무 크므로 결측값을 대체하기로 한다. knn 방법을 이용해 결측값을 채우되, 종속변수는 예측해서 채워 넣으면 예측력의 현실성이 떨어지므로 제외했다.

```
> library(DMwR)
> who = knnImputation(who[, !names(who)%in% c("Life.expendency","Status")], k=10)
> sum(is.na(who))
[1] 0
```

⇒ 따라서 결측값의 개수가 0이 됐음을 확인할 수 있다.

(4) 데이터 스케일링

이 데이터의 경우 데이터를 스케일링하면 설명변수의 영향을 직관적으로 해석하기 어려울 듯해서 굳이 표준화는 하지 않기로 한다. 측정 단위도 크게 차이나지 않는다.

3. 데이터 분석

(1) 다중선형회귀모형 적합

```
> fit = lm(Life.expectancy~., data=who)
> summary(fit)

Call:
lm(formula = Life.expectancy ~ ., data = who)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9826  -2.1480  -0.1211   2.2209  18.4179

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.260e+02  3.422e+01   3.682 0.000235 ***
Year                           -3.632e-02  1.711e-02  -2.123 0.033861 *
Adult.Mortality                -1.787e-02  7.774e-04 -22.990  < 2e-16 ***
infant.deaths                   9.421e-02  8.152e-03  11.557  < 2e-16 ***
Alcohol                         5.120e-02  2.437e-02   2.101 0.035734 *
percentage.expenditure          7.565e-05  8.563e-05   0.883 0.377045
Hepatitis.B                    -7.568e-03  4.074e-03  -1.858 0.063340 .
Measles                        -1.950e-05  7.385e-06  -2.641 0.008314 **
BMI                             3.488e-02  4.820e-03   7.237 5.85e-13 ***
under.five.deaths              -7.028e-02  5.979e-03 -11.755  < 2e-16 ***
Polio                           2.473e-02  4.360e-03   5.673 1.54e-08 ***
Total.expenditure               8.861e-02  3.276e-02   2.705 0.006874 **
Diphtheria                      3.154e-02  4.805e-03   6.565 6.16e-11 ***
HIV.AIDS                       -4.900e-01  1.705e-02 -28.745  < 2e-16 ***
GDP                             4.774e-05  1.319e-05   3.619 0.000301 ***
Population                     -5.611e-10  1.629e-09  -0.345 0.730485
thinness..1.19.years           -5.781e-02  4.865e-02  -1.188 0.234843
thinness.5.9.years             -9.203e-03  4.789e-02  -0.192 0.847629
Income.composition.of.resources 7.130e+00  6.201e-01  11.498  < 2e-16 ***
Schooling                       7.860e-01  4.176e-02  18.823  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.909 on 2916 degrees of freedom
Multiple R-squared:  0.8324,    Adjusted R-squared:  0.8313
F-statistic:   762 on 19 and 2916 DF,  p-value: < 2.2e-16
```

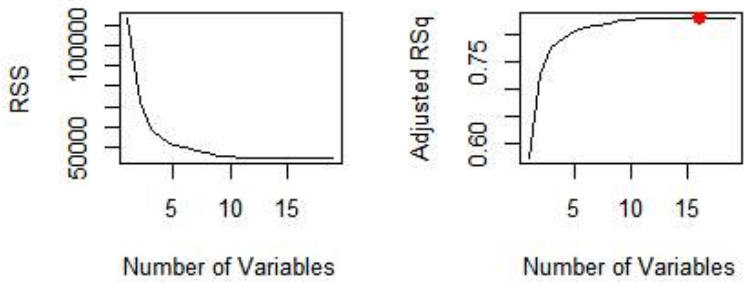
⇒ 제일 먼저 모든 변수를 가지고 다중선형회귀모형에 적합한 결과이다. 수정된 결정계수는 약 0.83이지만 유의하지 않은 변수가 많아 보인다. 설명력 없는 변수는 모형에 불필요한 복잡도를 높이므로 적절한 변수를 선택하기 위해 이어서 최적 부분집합 선택법과 계단선택법을 이용하겠다.

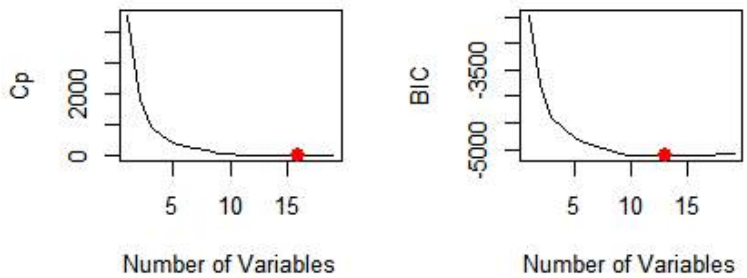
(2) 부분집합선택: 최적 부분집합 선택

```
> set.seed(0401)
> library(leaps)
> library(MASS)
> fit.best = regsubsets(Life.expectancy~.,nvmax=20,data=who)
> best.summary = summary(fit.best)
```

⇒ 먼저 최적 부분집합 선택법으로 모형을 적합했다. 그래프를 통해 변수 선택이 잘 되었는지 살펴보자

```
> par(mfrow=c(2,2))
> #1
> plot(best.summary$rss,xlab="Number of Variables",ylab="RSS",type="l") #RSS; Residual Sum of Squares
> #2
> plot(best.summary$adjr2,xlab="Number of Variables",ylab="Adjusted RSq",type="l") #Adjusted R^2
> which.max(best.summary$adjr2) #Adjr2의 최대값; 변수가 8개일 때
[1] 16
> points(16,best.summary$adjr2[16], col="red",cex=2,pch=20) #최대값 빨간점으로 찍기
> #3
> plot(best.summary$cp,xlab="Number of Variables",ylab="Cp",type='l') #Cp
> which.min(best.summary$cp) #Cp의 최소값
[1] 16
> points(16,best.summary$cp[16],col="red",cex=2,pch=20) #최소값 빨간점으로 찍기
> #4
> plot(best.summary$bic,xlab="Number of Variables",ylab="BIC",type='l') #BIC
> which.min(best.summary$bic) #BIC의 최소값
[1] 13
> points(13,best.summary$bic[13],col="red",cex=2,pch=20) #최소값 빨간점으로 찍기
```



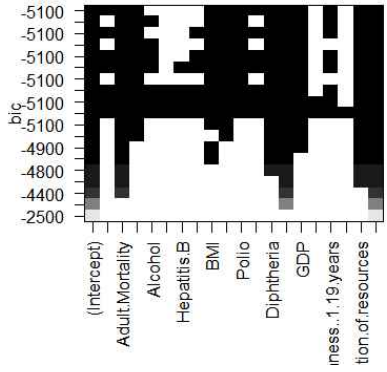


⇒ **모형선택 기준이 되는 RSS, Adjust R^2, C_p, BIC를 플롯한 결과이다. 각 빨간 점은 가장 최적의 변수의 개수를 나타낸다. 가장 적은 변수 개수를 가지는 BIC 기준으로 13개의 변수를 선택할 수 있다.**

```
> coef(fit.best,13)
                  (Intercept)                          Year               Adult.Mortality
                 1.507821e+02                 -4.890911e-02                 -1.759824e-02
                infant.deaths                          BMI               under.five.deaths
                 9.276511e-02                  3.464313e-02                 -6.974836e-02
                        Polio           Total.expenditure                     Diphtheria
                 2.334323e-02                  1.094772e-01                  2.764462e-02
                      HIV.AIDS                           GDP       thinness..1.19.years
                -4.892430e-01                  6.053389e-05                 -7.600370e-02
Income.composition.of.resources                    Schooling
                 7.285507e+00                  8.105323e-01
```

⇒ **선택한 13개의 변수와 추정된 회귀계수이다.**

```
> par(mfrow=c(1,1))
> plot(fit.best,scale="bic")
```



⇒ **또한 BIC를 이용해 적합한 결과를 계수 plot을 통해 확인할 수 있다.**

```
> best.summary$rsq[13]
[1] 0.8314151
```

⇒ **그리고 그 때의 결정계수는 약 0.83으로 모든 변수를 이용해 적합했을 때와 크게 다르지 않다. 즉, 변수를 더 조금 사용하고도 모형의 설명력을 잃지 않은 결과를 얻었다.**

부분집합 선택 방법 중 최적 부분집합 선택은 설명변수의 2^p개의 모든 가능한 모든 조합에서 최적 모형을 선택하는 방법이다. 따라서 p가 클수록 계산이 어려워져 비효율적이다. 따라서 p가 클 경우 전진선택법, 후진선택법, 계단선택법 등 다른 방법을 이용하는 게 더 좋을 수 있다. 특히 전진선택법과 후진선택법이 섞인 계단선택법을 이어서 살펴보겠다.

(3) 부분집합선택: 계단선택법

```
> fit = lm(Life.expectancy~., data=who)
> fit.con = lm(Life.expectancy~1, data=who)
> fit.step = step(fit.con,scope=list(lower=fit.con,upper=fit),direction="both")
Start:  AIC=13229.86
Life.expectancy ~ 1
> summary(fit.step)

Call:
lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + Diphtheria + BMI + GDP +
    Polio + Measles + thinness..1.19.years + Year + Total.expenditure +
    Hepatitis.B + under.five.deaths + infant.deaths + Alcohol,
    data = who)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0135  -2.1495  -0.1286   2.2216  18.4020

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.293e+02  3.402e+01   3.802 0.000147 ***
Schooling                        7.860e-01  4.172e-02  18.837  < 2e-16 ***
HIV.AIDS                        -4.903e-01  1.703e-02 -28.790  < 2e-16 ***
Adult.Mortality                 -1.787e-02  7.766e-04 -23.006  < 2e-16 ***
Income.composition.of.resources  7.107e+00  6.194e-01  11.474  < 2e-16 ***
Diphtheria                       3.152e-02  4.798e-03   6.570 5.93e-11 ***
BMI                              3.466e-02  4.774e-03   7.261 4.92e-13 ***
GDP                              5.806e-05  6.236e-06   9.311  < 2e-16 ***
Polio                            2.464e-02  4.356e-03   5.656 1.70e-08 ***
Measles                         -1.941e-05  7.375e-06  -2.632 0.008545 **
thinness..1.19.years            -6.646e-02  2.316e-02  -2.870 0.004135 **
Year                            -3.798e-02  1.701e-02  -2.233 0.025632 *
Total.expenditure                9.308e-02  3.240e-02   2.873 0.004094 **
Hepatitis.B                     -7.751e-03  4.059e-03  -1.909 0.056305 .
under.five.deaths               -7.007e-02  5.923e-03 -11.831  < 2e-16 ***
infant.deaths                    9.377e-02  8.007e-03  11.711  < 2e-16 ***
Alcohol                          5.268e-02  2.428e-02   2.170 0.030125 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.907 on 2919 degrees of freedom
Multiple R-squared:  0.8323,    Adjusted R-squared:  0.8314
F-statistic: 905.4 on 16 and 2919 DF,  p-value: < 2.2e-16
```

⇒ 계단선택법으로 선택된 최종 변수는 16개이며 수정된 결정계수는 약 0.83으로 최적 부분집합 선택법과 비슷한 듯하다. 역시 (1)에서 모든 변수를 사용한 경우와 비교해 변수가 줄었음에도 모형의 설명력은 비슷하다.

(4) 회귀진단

```
> fit = lm(Life.expectancy~Schooling+HIV.AIDS+Adult.Mortality+Income.composition.of.resources
+          +Diphtheria+BMI+GDP+Polio+Measles+thinness..1.19.years+Year+Total.expenditure
+          +Hepatitis.B+under.five.deaths+infant.deaths+Alcohol, data=who)
```

⇒ 위의 계단선택법으로 얻은 모형이 적절한지 알아보기 위해 회귀진단 해보겠다.

```
> library(car)
필요한 패키지를 로딩중입니다: carData
> vif(fit)
                     Schooling                       HIV.AIDS                Adult.Mortality
                      3.735261                       1.438482                       1.787049
Income.composition.of.resources                    Diphtheria                            BMI
                      3.224984                       2.481029                       1.760670
                           GDP                         Polio                       Measles
                      1.319423                       1.992359                       1.376004
          thinness..1.19.years                           Year            Total.expenditure
                      2.008875                       1.184055                       1.184273
                  Hepatitis.B              under.five.deaths                   infant.deaths
                      1.899236                     173.724474                     171.506063
                      Alcohol
                      1.793996
```

⇒ 먼저 다중공선성이 있는 변수를 확인하기 위해 VIF값을 확인해 보았다. 보통 VIF가 10을 넘는 변수는 다중공선성을 의심할 수 있다. 특히 under.five.deaths, infant.deaths의 VIF 값이 아주 크기 때문에 under.five.deaths 변수를 빼고 다시 적합하겠다.

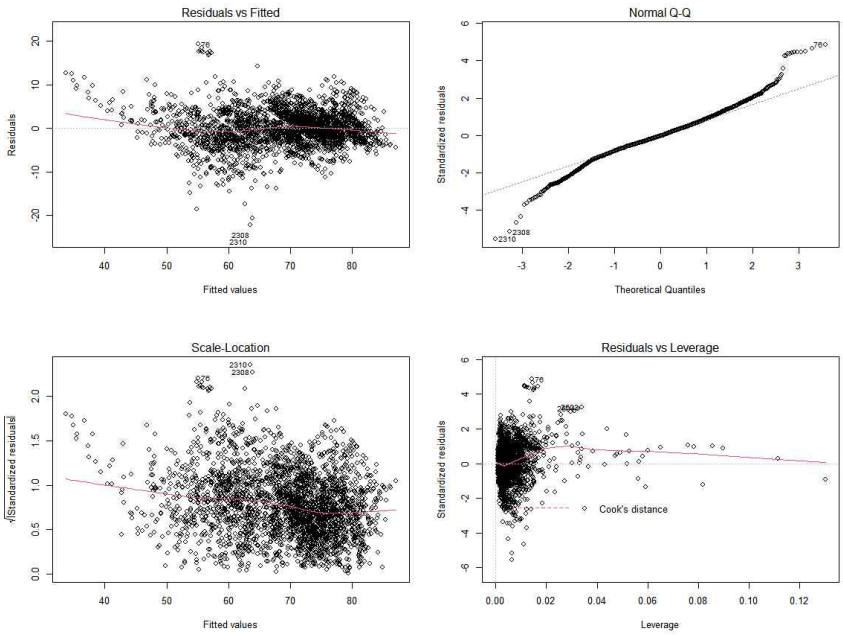
```
> fit2 = lm(Life.expectancy~Schooling+HIV.AIDS+Adult.Mortality+Income.composition.of.resources
+         +Diphtheria+BMI+GDP+Polio+Measles+thinness..1.19.years+Year+Total.expenditure
+         +Hepatitis.B+infant.deaths+Alcohol, data=who)
> vif(fit2)
                     Schooling                       HIV.AIDS                Adult.Mortality
                      3.720332                       1.434235                       1.785036
Income.composition.of.resources                    Diphtheria                            BMI
                      3.191933                       2.451471                       1.760600
                           GDP                         Polio                       Measles
                      1.315526                       1.986481                       1.362874
          thinness..1.19.years                           Year            Total.expenditure
                      2.007707                       1.181767                       1.184192
                  Hepatitis.B                   infant.deaths                        Alcohol
                      1.898183                       1.701553                       1.744934
```

⇒ 이제 VIF가 10을 넘는 변수가 없으므로 이 모형을 선택하고 나머지 가정에 대해서도 회귀진단을 하겠다.

```
> summary(fit2)

Call:
lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + Diphtheria + BMI + GDP +
    Polio + Measles + thinness..1.19.years + Year + Total.expenditure +
    Hepatitis.B + infant.deaths + Alcohol, data = who)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0940  -2.2097  -0.1243   2.2596  19.2980

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.466e+02  3.477e+01   4.217 2.55e-05 ***
Schooling                       8.175e-01  4.261e-02  19.185  < 2e-16 ***
HIV.AIDS                       -5.009e-01  1.740e-02 -28.786  < 2e-16 ***
Adult.Mortality                -1.820e-02  7.941e-04 -22.916  < 2e-16 ***
Income.composition.of.resources 7.855e+00  6.306e-01  12.457  < 2e-16 ***
Diphtheria                      3.773e-02  4.881e-03   7.731 1.46e-14 ***
BMI                             3.490e-02  4.884e-03   7.145 1.13e-12 ***
GDP                             5.413e-05  6.371e-06   8.495  < 2e-16 ***
Polio                           2.743e-02  4.451e-03   6.163 8.09e-10 ***
Measles                        -2.791e-05  7.511e-06  -3.716 0.000206 ***
thinness..1.19.years           -6.035e-02  2.369e-02  -2.548 0.010889 *
Year                           -4.725e-02  1.738e-02  -2.718 0.006607 **
Total.expenditure               9.772e-02  3.313e-02   2.950 0.003208 **
Hepatitis.B                    -8.939e-03  4.153e-03  -2.152 0.031445 *
infant.deaths                  -4.821e-04  8.161e-04  -0.591 0.554723
Alcohol                         2.609e-03  2.440e-02   0.107 0.914851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.999 on 2922 degrees of freedom
Multiple R-squared:  0.8242,    Adjusted R-squared:  0.8233
F-statistic: 913.2 on 15 and 2922 DF,  p-value: < 2.2e-16
```

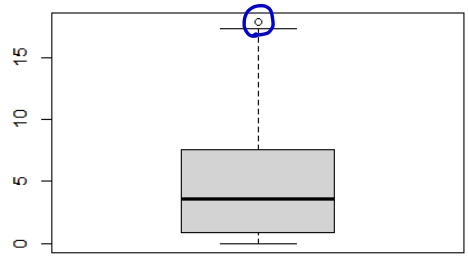
⇒ **수정된 결정계수: 0.8233**

```
> par(mfrow=c(2,2))
> plot(fit2)
```

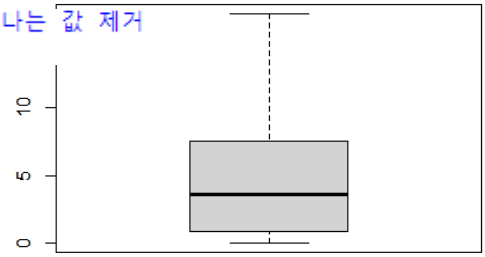


⇒ **선형성 가정과 등분산성 가정은 맞는 것 같으나 정규성 가정이 애매해 보인다. 이상치를 제거하고 다시 적합해보겠다.**

```
> boxplot(who$Alcohol) # 극단치 기준 확인
> boxplot(who$Alcohol)$stats #통계량 확인
      [,1]
[1,]  0.01
[2,]  0.85
[3,]  3.63
[4,]  7.56
[5,] 17.31
```




```
> who = who[who$Alcohol<17.31, ] # 기준을 벗어나는 값 제거
> boxplot(who$Alcohol) # 다시 확인
```



⇒ **변수 Alcohol의 이상치를 확인 후 제거까지 완료했다. 또한 잔차 그래프를 보면 76, 2308, 2310번째 관측이 이상점으로 판단되니 제거하고 모형에 적합하겠다.**

```
> fit3 = lm(Life.expectancy~Schooling+HIV.AIDS+Adult.Mortality+Income.composition.of.resources
+           +Diphtheria+BMI+GDP+Polio+Measles+thinness..1.19.years+Year+Total.expenditure
+           +Hepatitis.B+infant.deaths+Alcohol, data=who[-c(76,2308,2310),])
> summary(fit3)

Call:
lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + Diphtheria + BMI + GDP +
    Polio + Measles + thinness..1.19.years + Year + Total.expenditure +
    Hepatitis.B + infant.deaths + Alcohol, data = who[-c(76,
    2308, 2310), ])

Residuals:
     Min       1Q   Median       3Q      Max
-18.4828  -2.2190  -0.1446   2.2762  18.6474

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.530e+02  3.431e+01   4.458 8.57e-06 ***
Schooling                       8.249e-01  4.216e-02  19.568  < 2e-16 ***
HIV.AIDS                       -5.000e-01  1.717e-02 -29.124  < 2e-16 ***
Adult.Mortality                -1.838e-02  7.849e-04 -23.414  < 2e-16 ***
Income.composition.of.resources 7.865e+00  6.224e-01  12.638  < 2e-16 ***
Diphtheria                      3.698e-02  4.816e-03   7.679 2.18e-14 ***
BMI                             3.467e-02  4.819e-03   7.194 7.94e-13 ***
GDP                             5.376e-05  6.286e-06   8.553  < 2e-16 ***
Polio                           2.664e-02  4.392e-03   6.065 1.49e-09 ***
Measles                        -2.808e-05  7.411e-06  -3.789 0.000154 ***
thinness..1.19.years           -5.193e-02  2.339e-02  -2.221 0.026452 *
Year                           -5.046e-02  1.715e-02  -2.942 0.003289 **
Total.expenditure               1.198e-01  3.279e-02   3.655 0.000262 ***
Hepatitis.B                    -8.471e-03  4.098e-03  -2.067 0.038799 *
infant.deaths                  -5.521e-04  8.052e-04  -0.686 0.492964
Alcohol                        -1.907e-03  2.411e-02  -0.079 0.936986
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.945 on 2919 degrees of freedom
Multiple R-squared:  0.8281,    Adjusted R-squared:  0.8272
F-statistic: 937.6 on 15 and 2919 DF,  p-value: < 2.2e-16
```

⇒ **수정된 결정계수: 0.8272. 즉, 이상치 제거 전보다 약간 증가한 값이다.**

마지막으로 모형이 잘 적합 되었는지 예측구간 및 예측확률을 확인하면서 회귀진단을 마무리하겠다.

```
> # 예측구간
> set.seed(1)
> pred=predict(fit3, newdata=who, interval="predict")
> pred=as.data.frame(pred)
> head(pred) #예측구간
       fit      lwr      upr
1 61.23930 53.46180 69.01680
2 62.32489 54.56547 70.08432
3 62.43234 54.67376 70.19091
4 62.44370 54.68479 70.20262
5 62.02853 54.27086 69.78620
6 61.78560 54.02444 69.54675
```

```
> pred = cbind(pred,who$Life.expectancy)
> head(pred) #실제값과 비교
       fit      lwr      upr who$Life.expectancy
1 61.23930 53.46180 69.01680                65.0
2 62.32489 54.56547 70.08432                59.9
3 62.43234 54.67376 70.19091                59.9
4 62.44370 54.68479 70.20262                59.5
5 62.02853 54.27086 69.78620                59.2
6 61.78560 54.02444 69.54675                58.8
```

⇒ 예측구간과 실제값과 비교한 결과이다. 한눈에 확인하기 위해서 예측 성공을 True, 실패를 False로 해서 비율을 살펴보자.

```
> tf = NA
> pred = cbind(pred,tf)
> pred$tf[pred$'who$Life.expectancy'>=pred$lwr & pred$'who$Life.expectancy'<=pred$upr] = T
> pred$tf[is.na(pred$tf)] = F
> head(pred)
       fit      lwr      upr who$Life.expectancy   tf
1 61.23930 53.46180 69.01680                65.0 TRUE
2 62.32489 54.56547 70.08432                59.9 TRUE
3 62.43234 54.67376 70.19091                59.9 TRUE
4 62.44370 54.68479 70.20262                59.5 TRUE
5 62.02853 54.27086 69.78620                59.2 TRUE
6 61.78560 54.02444 69.54675                58.8 TRUE
> sum(pred$tf=="TRUE")/dim(pred)[1] #예측에 성공하는 비율이 약 93%로 성공적인 회귀분석 완료
[1] 0.9424779
```

⇒ 예측에 성공하는 비율이 약 94%로 성공적인 회귀분석을 마무리 하겠다.

이어서 다른 방법들로도 변수를 선택해 모형을 적합해보겠다. 특히 이번에는 데이터 셋을 훈련데이터와 시험데이터로 나눠 모형의 예측력을 비교할 것이다. Bias를 약간 희생하고 분산을 줄이는 축소추정법의 Ridge, Lasso방법과 차원축소를 이용하는 PCR과 PLS를 새로 적합하고 (1)의 다중선형회귀모형과 함께 비교해보자.

```
> set.seed(0401)
> train.size = dim(who)[1]/2
> train = sample(1:dim(who)[1], train.size)
> test = -train
> who.train = who[train, ]
> who.test = who[test, ]
```

⇒ 본격적인 모형 적합에 들어가기에 앞서 훈련데이터와 시험데이터를 각각 반으로 나눴다.

```
> lm.fit = lm(Life.expectancy~., data=who.train)
> lm.pred = predict(lm.fit, who.test)
> mean((lm.pred-who.test$Life.expectancy)^2)
[1] 16.21941
```

⇒ 다중선형회귀모형의 시험 mse: 16.2194

(5) 축소추정법: Ridge

```
> library(glmnet)
> train.X = model.matrix(Life.expectancy~., data=who.train)
> test.X = model.matrix(Life.expectancy~., data=who.test)
> grid = 10^seq(4,-2,length=100)
> ridge.fit = glmnet(train.X, who.train$Life.expectancy, alpha=0, lambda=grid, thresh=1e-12)
> ridge.cv = cv.glmnet(train.X, who.train$Life.expectancy, alpha=0, lambda=grid, thresh=1e-12)
> bestlambda.R = ridge.cv$lambda.min
> bestlambda.R
[1] 0.01
> #test error
> ridge.pred = predict(ridge.fit, s=bestlambda.R, newx=test.X)
> mean((ridge.pred - who.test$Life.expectancy)^2)
[1] 16.22193
```

⇒ **Ridge의 시험 mse: 16.22193, λ=0.01**

Ridge는 최소제곱 추정의 분산이 큰 경우 편의를 약간 희생하고 분산을 줄여 효율적인 추정을 하는 방법이다. 시험 mse는 비슷하게 나왔지만 부분집합 선택방법보다 계산이 적은 장점이 있다.

(6) 축소추정법: Lasso

```
> lasso.fit = glmnet(train.X, who.train$Life.expectancy, alpha=1, lambda=grid, thresh=1e-12)
> lasso.cv = cv.glmnet(train.X, who.train$Life.expectancy, alpha=1, lambda=grid, thresh=1e-12)
> bestlambda.L = lasso.cv$lambda.min
> bestlambda.L
[1] 0.01
> #test error
> lasso.pred = predict(lasso.fit, s=bestlambda.L, newx=test.X)
> mean((lasso.pred - who.test$Life.expectancy)^2)
[1] 16.22651
```

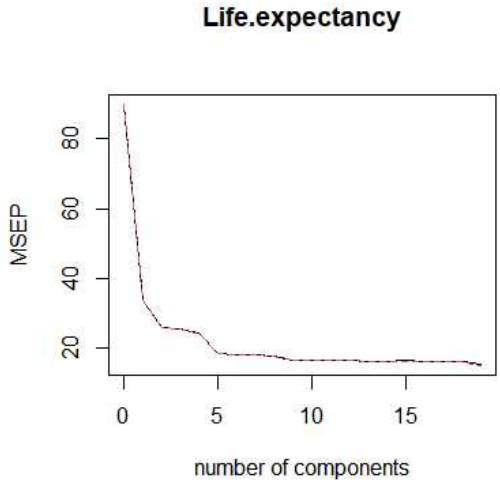
⇒ **Lasso의 시험 mse: 16.2265, λ=0.01**

```
> predict(lasso.fit, s=bestlambda.L, type="coefficients")
21 x 1 sparse Matrix of class "dgCMatrix"
                                       1
(Intercept)                 1.375314e+02
(Intercept)                 .
Year                       -4.163935e-02
Adult.Mortality            -1.914885e-02
infant.deaths               7.256337e-02
Alcohol                     1.861745e-02
percentage.expenditure      1.520837e-04
Hepatitis.B                -9.136883e-03
Measles                    -9.206207e-06
BMI                         3.700052e-02
under.five.deaths          -5.614324e-02
Polio                       2.428926e-02
Total.expenditure           2.075108e-02
Diphtheria                  3.576200e-02
HIV.AIDS                   -4.997399e-01
GDP                         3.246485e-05
Population                  1.982052e-09
thinness..1.19.years       -4.476147e-02
thinness.5.9.years         -3.863618e-02
Income.composition.of.resources  6.771660e+00
Schooling                   7.848260e-01
```

⇒ **Lasso는 Ridge와는 달리 변수선택을 하므로 일부 회귀계수는 0이 되기도 한다. 그러나 이 경우 0이 되는 회귀계수는 없는 듯하다.**

(7) 차원축소: PCR

```
> library(pls)
> pcr.fit = pcr(Life.expectancy~., data=who.train, scale=TRUE, validation="CV")
> validationplot(pcr.fit, val.type="MSEP")
```



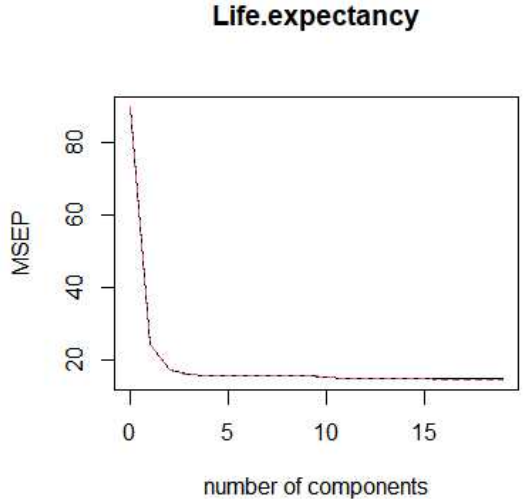
⇒ **변수가 1->2개, 2->3개가 될 때 cv가 가장 크게 감소한다. 변수가 10개 선택될 때 가장 좋아보인다.**

```
> pcr.pred = predict(pcr.fit, who.test, ncomp = 10)
> mean((pcr.pred - who.test$Life.expectancy)^2)
[1] 17.39058
```

⇒ **PCR의 시험 mse: 17.39058, 시험 오차는 지금까지 중 가장 크다.**

(8) 차원축소: PLS

```
> pls.fit = plsr(Life.expectancy~., data=who.train, scale=TRUE, validation="CV")
> validationplot(pls.fit, val.type="MSEP")
```



⇒ **역시 변수가 3개정도 까지 선택될 때 cv가 가장 크게 감소한다. 변수가 11개 선택될 때 가장 좋아보인다.**

```
> pls.pred = predict(pls.fit, who.test, ncomp = 11)
> mean((pls.pred - who.test$Life.expectancy)^2)
[1] 16.38829
```

⇒ **PLS의 시험 mse: 16.38829**

(9) 모형비교

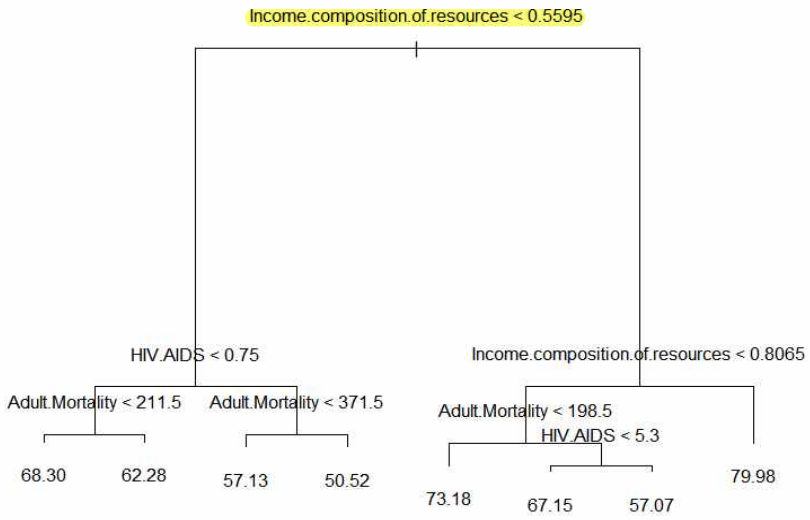
```
> test.avg = mean(who.test$Life.expectancy)
> lm.r2 = 1-mean((lm.pred - who.test$Life.expectancy)^2)/mean((test.avg - who.test$Life.expectancy)^2)
> ridge.r2 = 1-mean((ridge.pred - who.test$Life.expectancy)^2) / mean((test.avg - who.test$Life.expectancy)^2)
> lasso.r2 = 1-mean((lasso.pred - who.test$Life.expectancy)^2) / mean((test.avg - who.test$Life.expectancy)^2)
> pcr.r2 = 1-mean((pcr.pred - who.test$Life.expectancy)^2) / mean((test.avg - who.test$Life.expectancy)^2)
> pls.r2 = 1-mean((pls.pred - who.test$Life.expectancy)^2) / mean((test.avg - who.test$Life.expectancy)^2)
> all.r2 = c(lm.r2, ridge.r2, lasso.r2, pcr.r2, pls.r2)
> all.r2
[1] 0.8224836 0.8224561 0.8224059 0.8096655 0.8206353
```

⇒ **지금까지 적합한 모형의 결정계수 값을 확인해보니 PCR을 제외한 나머지 방법들의 모형 설명력이 비슷함을 알 수 있다. 이는 시험 mse를 봐도 비슷한 결과이다.**

모형 적합 결과 시험 MSE와 결정계수 관점 모두 PCR을 제외하고는 다 비슷한 성능을 가진 듯하다. 이제 지금까지의 모형 적합을 바탕으로 기대수명과 설명변수의 관계를 설명할 것이다. 다만, 쉬운 해석을 위해 마지막으로 회귀나무와 랜덤포레스트를 통해 가장 영향이 큰 변수만 확인하고 분석을 마무리 하겠다.

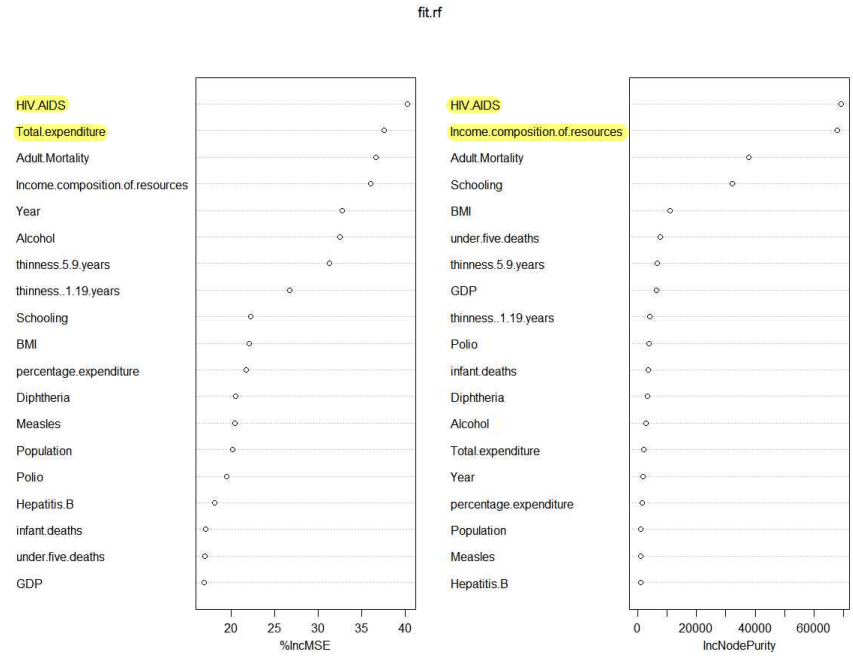
(10) 회귀나무 및 랜덤포레스트

```
> library(tree)
> tree.who = tree(Life.expectancy~., data=who.train)
> summary(tree.who)

Regression tree:
tree(formula = Life.expectancy ~ ., data = who.train)
Variables actually used in tree construction:
[1] "Income.composition.of.resources" "HIV.AIDS"                        "Adult.Mortality"
Number of terminal nodes:  8
```



⇒ **가장 영향력이 큰 변수와 나눠지는 기준을 분기점을 통해 쉽게 확인할 수 있다.**

```
> library(randomForest)
> fit.rf = randomForest(Life.expectancy~., data=who, importance=TRUE)
> varImpPlot(fit.rf)
```



⇒ 또는 랜덤포레스트를 통해서도 가장 중요한 변수를 확인할 수 있다.

4. 데이터 분석 결과

먼저 분석(11)의 결과를 보면 기대수명에 가장 큰 영향을 미치는 설명변수는 "HIV.AIDS"와 "Income.composition.of.resources"이다. 이때 HIV.AIDS 변수는 생아 1000명당 HIV바이러스에 대한 사망률이며 Income- 변수는 인간개발지수를 나타낸다. 이를 분석(4)의 결과 최종적으로 얻은 모형의 계수를 통해 해석하면 다음과 같다.

(i) 생아(0-4세) 1,000명당 HIV/AIDS에 대한 사망률이 높을수록 기대수명이 짧다.
(ii) 인간개발지수가 높을수록 기대수명이 길다

이는 상식적으로 생각해도 예측할 수 있는 당연한 결과인 듯하다. 대신 이 분석을 통해 그런 변수들의 중요도를 매겼을 때 인간개발지수(Income-)와 HIV 바이러스(HIV.AIDS)가 음주(Alcohol)와 비만도(BMI)보다 중요하다는 새로운 사실을 알 수 있었다. 또한 특이한 점은 몇몇 바이러스와 기대수명이 양의 상관관계를 가지는 걸로 해석된다. 마지막으로 유아사망률(infant deaths)과 음주(Alcohol)는 기대수명과 음의 상관관계를 가지지만 유의확률이 크기 때문에 통계적으로 유의하다고 하기는 어렵다.

3. 결론

1. 결과요약

(1) 기대수명과 양의 상관관계를 가지는 요인:
학업 수준, 인간개발지수, 지출 수준, GDP 등
(2) 기대수명과 음의 상관관계를 가지는 요인:
HIV 바이러스, 홍역 바이러스, 10대의 thinness(심하게 여윔), 성인 사망률 등
(3) 기대수명에 영향력이 큰 요인:
인간개발지수≈HIV바이러스>성인사망률>학업 수준 등

2. 활용방안

기대수명은 우리 삶과 밀접한 관계가 있는 요소로 특히 의학 통계 분야 등에서 활용될 수 있다. 이번 분석을 통해 기대수명에 영향을 미치는 변수를 찾고 그 변수의 영향의 정도를 알 수 있었다. 이제 또 다른 분석을 통해 또 다른 분석을 통해 어떤 변수를 조정할 때 최소 비용으로 최대로 기대수명을 늘릴 수 있는지 등을 비교해 어떤 변수에 투자하는 것이 효율적인지의 연구로 이어질 수 있다. 기대수명이 증가한다는 것은 단지 오래 살 수 있다는 의미가 아니라 그만큼 건강 수준이 높아져 인구의 생산성 향상을 이끌어 국가의 경제 발전으로도 이어지므로 중요한 연구 중 하나이다.