#0.0 원시 데이터셋

```
setwd("C:/Users/DS/Downloads")
maindata = read.csv("maindata.csv")
dim(maindata) # 6265 623
```

#0.1 사용할 변수만 있는 데이터셋

```
#변수이름 데이터셋
name = read.csv("name.csv", header=T)

str(name)
name

data = maindata[, name[,1]]
#write.csv(data , file = "C:/Users/KJPark/Downloads/data2.csv", row.names = F)
```

#0.2 NA행 제거

```
#data = read.csv("data2.csv")

#사용자 함수
library(dplyr)
fil <- function(df, variables){
  new=df
  for (variable in variables){
    new <- new %>% filter(!is.na(new[[variable]]))

  }
  return(new)
}

data.no.na = fil(data, name[-3,1])
dim(data.no.na)
#write.csv(data.no.na , file = "C:/Users/KJPark/Downloads/data.no.na.csv",
row.names = F)
```

0.3.0 비해당(청소년) 행 제거 처리 전 데이터

```
data=data.no.na
data = read.csv("data.no.na.csv")

f.name=read.csv("f.name.csv",header=T)
data[,f.name$name]

to.factors <- function(df, variables){
  for (variable in variables){
```

```
    df[[variable]] <- as.factor(df[[variable]])
  }
  return(df)
}

data=to.factors(data, f.name$name)
str(data)
dim(data) #5030 79

# 0.3.1 변수 생성

# 0.3.1.1 반응변수
#고혈압 여부 DI1_dg
# 0.없음 | 1.있음 | 8.비해당
data$DI1_dg
sum(is.na(data$DI1_dg))
table(data$DI1_dg)
barplot(table(data$DI1_dg),main="고혈압없음(0) vs 고혈압있음(1) vs비해당(8)")
#==================================
#이상지질혈증 여부 DI2_dg
# 0.없음 | 1.있음 | 8.비해당
data$DI2_dg
sum(is.na(data$DI2_dg))
table(data$DI2_dg)
barplot(table(data$DI2_dg),main="이상지질혈증 여부: 없음(0) vs 있음(1) vs비해당(8)")
#==================================
#당뇨병 여부 DE1_dg
# 0.없음 | 1.있음 | 8.비해당
data$DE1_dg
sum(is.na(data$DE1_dg))
table(data$DE1_dg)
barplot(table(data$DE1_dg),main="당뇨병 여부: 없음(0) vs 있음(1) vs비해당(8)")
#==================================
#천식 여부 DJ4_dg
# 0.없음 | 1.있음 | 8.비해당 | 9.무응답
data$DJ4_dg
sum(is.na(data$DJ4_dg))
table(data$DJ4_dg)
barplot(table(data$DJ4_dg),main="천식 여부: 없음(0) vs 있음(1) vs비해당(8) vs 무응답(9)")
#무응답 처리 필요
#==================================
#아토피 여부 DL1_dg
# 0.없음 | 1.있음 | 8.비해당 | 9.무응답
data$DL1_dg
sum(is.na(data$DL1_dg))
table(data$DL1_dg)
```

```
barplot(table(data$DL1_dg),main="아토피 여부: 없음(0) vs 있음(1) vs 비해당(8) vs
무응답(9)")
#무응답 처리 필요
#==================================
#알레르기비염 여부 DJ8_dg
# 0.없음 | 1.있음 | 8.비해당 | 9.무응답
data$DJ8_dg
sum(is.na(data$DJ8_dg))
table(data$DJ8_dg)
barplot(table(data$DJ8_dg),main="알르레기 비염 여부: 없음(0) vs 있음(1) vs 비해당(8) vs
무응답(9)")
# 무응답 처리 필요
#==================================
#부비동염 여부 DJ6_dg
# 0.없음 | 1.있음 | 8.비해당 | 9.무응답
data$DJ6_dg
sum(is.na(data$DJ6_dg))
table(data$DJ6_dg)
barplot(table(data$DJ6_dg),main="부비동염 여부: 없음(0) vs 있음(1) vs 비해당(8) vs
무응답(9)")
# 무응답 처리 필요
#==================================
#중이염 여부 DH4_dg
# 0.없음 | 1.있음 | 8.비해당 | 9.무응답
data$DH4_dg
sum(is.na(data$DH4_dg))
table(data$DH4_dg)
barplot(table(data$DH4_dg),main="중이염 여부: 없음(0) vs 있음(1) vs 비해당(8) vs
무응답(9)")
# 무응답 처리 필요
#==================================
#콩팥병 여부 DN1_dg
# 0.없음 | 1.있음 | 8.비해당 | 9.무응답
data$DN1_dg
sum(is.na(data$DN1_dg))
table(data$DN1_dg)
barplot(table(data$DN1_dg),main="콩팥병 여부: 없음(0) vs 있음(1) vs 비해당(8) vs
무응답(9)")
# 무응답 처리 필요, 비해당(청소년, 소아) 데이터가 392개로 같은 수임
```

#0.3.1.2 설명변수

```
#==================================
#수도권 city
#1. 수도권(1=서울, 4=인천, 9=경기), 0. 비수도권(그 외)
data$region
sum(is.na(data$region))
## 결측치 없음
```

```
table(data$region)
barplot(table(data$region), main = "region")

# 0.비수도권 | 1.수도권
data$city=c()
for(i in 1:dim(data)[1]){
  if (data$region[i]==1 | data$region[i]==4 | data$region[i]==9){
    data$city[i]= 1} #수도권
  else data$city[i]= 0
}
data$city
table(data$city)
barplot(table(data$city), main = "비수도권0 vs. 수도권1")

#region 변수제거
#city 변수 추가
#================================================
#남성, 임신 안한 여성, 임신한 여성 sex3
table(data$sex)
barplot(table(data$sex), main = "sex")

data$sex3=c()
for(i in 1:dim(data)[1]){
  if (data$sex[i]==1){
    data$ sex3[i]= 0} #남자
  else if (data$sex[i]==2 && data$LW_pr[i]==1){
    data$sex3[i]= 1} #여자이면서 임신
  else if (data$sex[i]==2 & data$LW_pr[i]==2){
    data$sex3[i]= 2} #여자이면서 임신 no
  else if (data$sex[i]==2 & data$LW_pr[i]==8){
    data$sex3[i]= 2} #여자 청소년 임신no
  else if (data$sex[i]==2 & data$LW_pr[i]==9){
    data$sex3[i]= 9} #여자인데 무응답
}

data$sex3
sum(!is.na(data$sex3))
table(data$sex)
table(data$sex3)

barplot(table(data$sex3),main = "남자(0) vs. 여자&임신O(1) vs. 여자&임신X(2)")
#무응답(9) 처리 필요: 여자임에도 임신경험여부 무응답
#sex, LW_pr 변수제거
#sex3 변수추가

#==================================
#나이 age
```

```
sum(is.na(data$age))
hist(data$age)
#==================================
#소득 ainc
data$ainc
sum(is.na(data$ainc))
hist(data$ainc)
#==================================
#결혼상태 marri_2
sum(is.na(data$marri_2))
table(data$marri_2)

barplot(table(data$marri_2))
# 1:유배우자,동거 | 2.유배우자,별거 | 3.사별
# 4.이혼 | 88. 비해당(결혼X) | 99. 무응답

#무응답과 비해당, 별거와 이혼
# 1:유배우자,동거 | 2.별거 및 이혼 | 3.사별
# 4. 비해당(결혼X) | 99.무응답

data$marri_2=factor(data$marri_2, levels=c(1,2,3,4,88,99), labels=c(1,2,3,2,4,99))
table(data$marri_2)

barplot(table(data$marri_2),main = "결혼(1) vs 별거(2) vs 사별(3) vs 결혼x(4)")
#4 결혼X와 비해당 >> 다른 변수를 통해 청소년 행 제거 시 결혼x 데이터만 남음
#99 무응답 2개 처리필요
#==================================
#민간의료보험가입여부 npins
#1.예 | 2.아니요 | 9.모름,무응답
sum(is.na(data$npins))
table(data$npins)
barplot(table(data$npins))
# 무응답 처리 필요(무응답 28개)
#==================================
#건강검진여부 BH1
#1.예 | 2.아니요 |8.청소년 | 9.모름,무응답
sum(is.na(data$BH1))
table(data$BH1)
barplot(table(data$BH1))

#0.아니요 | 1.예  |8.청소년 | 9.모름,무응답
data$BH1=factor(data$BH1, levels=c(1,2,8,9), labels=c(1,0,8,9))
table(data$BH1)
barplot(table(data$BH1),main = "건강검진여부: 아니요(0) vs 예(1) vs 청소년(8) vs 모름(9)")
#무응답 처리 필요
#==================================
#근로시간 EC_wht_23
#888: 근로X 999:무응답
```

```
sum(is.na(data$EC_wht_23))
sum(data$EC_wht_23==0)
sum(data$EC_wht_23==888)
sum(data$EC_wht_23==999)

for(i in 1:dim(data)[1]){
  if (data$EC_wht_23[i]==888){
    data$EC_wht_23[i]= 0}
}
sum(data$EC_wht_23==888)
sum(data$EC_wht_23==0)
sum(data$EC_wht_23==999)

hist(data$EC_wht_23)

#비해당 청소년과 근로하지 않은 성인이 모두 0으로 입력
#후에 청소년 행 제거시 자동으로 근로시간 0인 성인만 남음
#==================================
#체중변화량 weight.ch
#(기존) 체중변화 BO1_1
#1.변화X | 2.체중감소 | 3.체중증가 | 8.청소년 | 9.모름
table(data$BO1_1)

#체중감소량 BO1_2
#1. 3~6 | 2.6~10 | 3. 10이상 | 8.해당X | 9.모름
#체중감소 729
table(data$BO1_1) #2체중감소
table(data$BO1_2)
#1,2,3 이 감소한 경우 558  111   59   >>728
#한명이 얼마나 감소한지 몰라서 9응답

#체중증가량 BO1_3
#1. 3~6 | 2.6~10 | 3. 10이상 | 8.해당X | 9.모름
#체중증가 1108
table(data$BO1_1) #3 체중증가
table(data$BO1_3)
#1,2,3 이 증가한 경우 833  201   74  >>1108 동일

data$weight.ch =c()
for(i in 1:dim(data)[1]){
  if (data$BO1_1[i]==1){
    data$weight.ch[i]= 0 }  #변화X
  else if (data$BO1_1[i] == 2){ #체중감소
    data$weight.ch[i]= data$BO1_2[i] * (-1)}
  else if (data$BO1_1[i] == 3){ #체중증가
    data$weight.ch[i]= data$BO1_3[i] }
  else data$weight.ch[i]= data$BO1_1[i]
}
```

```
table(data$weight.ch)

#청소년8, 무응답(-9와 9)
#BO1_1,BO1_2,BO1_3 변수제거
#weight.ch 변수추가
#==================================
# 음주량 BD.total
#(기존) 음주량 BD1
# 평생 음주경험
# 1. 술을 마셔 본 적 없음 | 2. 있음 | 8. 비해당(소아) | 9. 모름/무응답
table(data$BD1)

#음주빈도 BD1_11
# 1: 전혀X | 2.월1회미만 | 3.월1회 | 4.월2~4회
# 5. 주 2~3회 | 6.주4회이상
# 8.비해당 | 9.모름

sum(is.na(data$BD1_11))
table(data$BD1_11)

data$BD1_11.new = c()
for(i in 1:dim(data)[1]){
  if (data$BD1_11[i]==1){
    data$BD1_11.new[i]= 0 }
  else if (data$BD1_11[i] == 2){ #
    data$BD1_11.new[i]= 0.5}
  else if (data$BD1_11[i] == 3){
    data$BD1_11.new[i]= 1 }
  else if (data$BD1_11[i] == 4){
    data$BD1_11.new[i]= 3 }
  else if (data$BD1_11[i] == 5){
    data$BD1_11.new[i]= 8 }
  else if (data$BD1_11[i] == 6){
    data$BD1_11.new[i]= 16 }   else if (data$BD1_11[i] == 8 & data$BD1[i] == 1){
      data$BD1_11.new[i]= 0 } # 평생 술 안 마심 & 현재 비해당
  else if (data$BD1_11[i] == 8 & data$BD1[i] == 8){
    data$BD1_11.new[i]= 88 } # 소아청소년
  else if (data$BD1_11[i] == 9){
    data$BD1_11.new[i]= 99 } # 무응답
}

table(data$BD1_11.new)
# 0 술 한번도 안 마심, 0.5~16 술 마시는 횟수, 88 소아청소년, 99 무응답

#한번에 마시는 음주량 BD2_1
table(data$BD2_1)
#1. 1~2잔 | 2. 3~4잔 | 3. 5~6잔 | 4. 7~9잔
#5. 10잔 이상 | 8. 비해당(술 아예 안 마심, 소아청소년) | 9. 무응답
```

```
data$BD2_1.new = c()
for(i in 1:dim(data)[1]){
  if (data$BD2_1[i]==1){
    data$BD2_1.new[i]= 1.5 }
  else if (data$BD2_1[i] == 2){ #
    data$BD2_1.new[i]= 3.5}
  else if (data$BD2_1[i] == 3){
    data$BD2_1.new[i]= 5.5 }
  else if (data$BD2_1[i] == 4){
    data$BD2_1.new[i]= 8 }
  else if (data$BD2_1[i] == 5){
    data$BD2_1.new[i]= data$BD2_14[i] }

  else if (data$BD2_1[i] == 8 & data$BD1_11.new[i] == 0){
    data$BD2_1.new[i]= 0 }
  #음주량 비해당, 음주빈도 0회+평생 술 안 마심 #성인인데 술안마심
  else if (data$BD2_1[i] == 8 & data$BD1_11[i] == 8){
    data$BD2_1.new[i]= 88 }
  #음주량 비해당, 음주빈도는 비해당 #청소년이라서 안마심
  else if (data$BD2_1[i] == 9){
    data$BD2_1.new[i]= 99 }
}

table(data$BD2_1.new)
hist(data$BD2_1.new)

#data$BD.total 음주량 최종변수 생성
table(data$BD1_11.new)
table(data$BD2_1.new)

data$BD.total <- c()
for (i in 1:dim(data)[1]){
  if(data$BD1_11.new[i] == 88) {data$BD.total[i] <- 888} #소아청소년 888
  else if(data$BD1_11.new[i] == 99) {data$BD.total[i] <- 999} #무응답 999
  else if(data$BD1_11.new[i] <= 16) {data$BD.total[i] <-
data$BD1_11.new[i]*data$BD2_1.new[i]} #음주량
}
table(data$BD.total)
#데이터 범위 0~400, 소아청소년(888) 114개 무응답(999) 23개

sum(data$BD.total == 888)
sum(data$BD.total == 999)

hist(data$BD.total)
hist(data$BD.total[data$BD.total< 888])
# 소아청소년(888). 무응답(999)
#BD1, BD1_11,BD1_11.new,BD2_1,BD2_1.new 변수제거
```

```
#BD.total 변수추가

#변수log 변환
for(i in 1:dim(data)[1]){
  if(data$BD.total[i] < 888){data$BD.total[i]= log( data$BD.total[i] + 1) }
  else if(data$BD.total[i] == 888) {data$BD.total[i] = 888}
  else if(data$BD.total[i] == 999) {data$BD.total[i] = 999}
}

hist(data$BD.total[data$BD.total< 888])

#==================================
#일주일 수면시간  sleep
#일주일 수면시간 주중:BP16_1 주말:BP16_2
#88 소아 청소년 99모름 무응답
sum(is.na(data$BP16_1))
sum(data$BP16_1==0)
sum(data$BP16_1==88)
sum(data$BP16_1==99)

hist(data$BP16_1)
hist(data[data$BP16_1 < 80, "BP16_1"])
mean( data[data$BP16_1 < 80, "BP16_1"] )

# 88, 99 행 겹치는지 확인
sum(data$BP16_1==88)
sum(data$BP16_2==88)
sum(data$BP16_1==88 & data$BP16_2==88 )

sum(data$BP16_1==99)
sum(data$BP16_2==99)
sum(data$BP16_1==99 & data$BP16_2==99 )

summary(data$BP16_1[data$BP16_1<=87])
summary(data$BP16_2[data$BP16_1<=87])

data$sleep=c()
for(i in 1:dim(data)[1]){
  if (data$BP16_1[i]<80){
    data$sleep[i]= data$BP16_1[i] * 5 + data$BP16_2[i] *2 }  #변화X
  else if(data$BP16_1[i] == 88) {data$sleep[i] = 888}
  else data$sleep[i] =999
}

sum(data$BP16_1==88)
sum(data$sleep==888)
sum(data$BP16_1==99)
sum(data$sleep==999)
```

```
# 소아청소년 데이터 888, 무응답 999로 코딩

hist(data[data$sleep < 888, "sleep"])
#소아 청소년,무응답 처리 필요

#BP16_1, BP16_2 변수제거
#sleep 변수추가
#==================================
# 스트레스 stress
sum(is.na(data$BP1))
table(data$BP1)

# 스트레스 인지정도 역코딩
# 1. 거의 안 느낌 | 2. 조금 느낌 | 3. 많이 느낌 | 4. 대단히 많이 느낌 | 8. 비해당 | 9. 무응답
# 숫자 커질수록, 스트레스 커지는 형태로 역코딩 진행

data$stress <- c()
for (i in 1:dim(data)[1]){
  if(data$BP1[i] <= 4) {data$stress[i] = 5-data$BP1[i]}
  else {data$stress[i] <- data$BP1[i]}
}
table(data$stress)

#install.packages("tidyverse")
library(tidyverse)

barplot(table(data$stress),main="거의 안 느낌(1) vs 조금느낌(2) vs많이 느낌(3) vs 대단히 많이 느낌(4) vs비해당(8) vs무응답(9)")

# BP1제거, stress 변수추가
# 무응답 처리 필요
#==================================
#흡연량 smoke_sum
table(data$BS1_1)
# 1. 5값(100개비) 미만 | 2. 5값 이상 | 3. 피운적x | 8. 비해당(소아청소년) | 9. 무응답

table(data$BS3_1)
# 1. 5값(100개비) 미만 | 2. 5값 이상 | 3. 피운적x | 8. 비해당(소아청소년, 담배x) | 9. 무응답

table(data$BS3_2)
# BS3_1의 비해당(담배x, 소아청소년)3163과 BS3_1의 (3.피운적x)1134가 합해져 4297
# 평생 흡연x. 과거 흡연했지만 이제 안함은 0 | 비해당 888 | 무응답 999로 새롭게 코딩

table(data$BS12_37)
#1. 흡연함 | 2. 흡연x | 8. 비해당(소아청소년) | 9. 무응답

table(data$BS12_47)
#1. 매일 | 2. 가끔 | 3. 과거에 피웠으나, 현재x | 8. 비해당(소아청소년, 담배x) | 9. 무응답
```

```
table(data$BS12_47_1)
# BS12_47의 비해당(담배x, 소아청소년)4621과  BS12_47의 (3.피운적x)210가 합해져 4831
# 평생 흡연x. 과거 흡연했지만 이제 안함은 0 | 비해당 888 | 무응답 999로 새롭게 코딩
# 현재 일반담배 흡연여부부터 합쳐서 코딩
#1. 흡연함 | 2. 흡연 안함 |  8. 비해당(소아청소년) | 9. 무응답으로 코딩
data$smoke_ox <- c()
for (i in 1:dim(data)[1]){
  if(data$BS3_1[i] <= 2) {data$smoke_ox[i] <- 1} #현재 일반담배 흡연함
  else if(data$BS3_1[i] == 3) {data$smoke_ox[i] <- 0} #현재 일반담배 흡연 안함
  else if(data$BS3_1[i] == 8 & data$BS1_1[i] == 3) {data$smoke_ox[i] <- 0} #평생 담배X
  else if(data$BS3_1[i] == 8 & data$BS1_1[i] == 8) {data$smoke_ox[i] <- 8} #청소년
  else if(data$BS3_1[i] == 9) {data$smoke_ox[i] <- 9} #무응답
}

table(data$smoke_ox)

## 현재 궐련형 전자담배 흡연여부
#1. 흡연함 | 2. 흡연 안함 |  8. 비해당(소아청소년) | 9. 무응답으로 코딩
data$smoke_ox2 <- c()
for (i in 1:dim(data)[1]){
  if(data$BS12_47[i] <= 2) {data$smoke_ox2[i] <- 1} #현재 전자담배 흡연
  else if(data$BS12_47[i] == 3) {data$smoke_ox2[i] <- 0} #현재 전자담배 흡연X
  else if(data$BS12_47[i] == 8 & data$BS12_37[i] == 2) {data$smoke_ox2[i] <- 0} #평생
담배x
  else if(data$BS12_47[i] == 8 & data$BS12_37[i] == 8) {data$smoke_ox2[i] <- 8} #청소년
  else if(data$BS12_47[i] == 9) {data$smoke_ox2[i] <- 9} #무응답
}

table(data$smoke_ox2)

# 흡연여부 합치기(일반담배&전자담배)
#0. 흡연 안 함 | 2. 흡연(일반 or 전자담배 사용) |  8. 비해당(소아청소년) | 9. 무응답으로
코딩

data$smoke_ox3 <- c()
for (i in 1:dim(data)[1]){
  if(data$smoke_ox[i] == 0 & data$smoke_ox2[i] == 0) {data$smoke_ox3[i] <- 0}
#일반&전담=비흡연
  else if(data$smoke_ox[i] == 1 | data$smoke_ox2[i] == 1) {data$smoke_ox3[i] <- 1} #일반 or
전담 흡연
  else if(data$smoke_ox[i] == 8) {data$smoke_ox3[i] <- 8} #청소년
  else if(data$smoke_ox[i] == 9) {data$smoke_ox3[i] <- 9} #무응답
}

table(data$smoke_ox3)

## 흡연여부 바탕으로, 흡연개수 연속형으로 변환
```

```
data$smoke_sum <- c()
for (i in 1:dim(data)[1]){
  if(data$smoke_ox3[i]==0) {data$smoke_sum[i] <- 0} #흡연 안함=0
  else if(data$smoke_ox[i] == 1 & data$smoke_ox2[i] == 1){data$smoke_sum[i] <-
data$BS3_2[i] + data$BS12_47_1[i]} #일반&전자담배 둘 다O
  else if(data$smoke_ox[i] == 1 & data$smoke_ox2[i] == 0){data$smoke_sum[i] <-
data$BS3_2[i] +0}#일반담배만
  else if(data$smoke_ox[i] == 0 & data$smoke_ox2[i] == 1){data$smoke_sum[i] <- 0 +
data$BS12_47_1[i]} #전자담배만
  else if(data$smoke_ox3[i] == 8) {data$smoke_sum[i] <- 888} #청소년
  else if(data$smoke_ox3[i] == 9) {data$smoke_sum[i] <- 999} #무응답
}

table(data$smoke_sum)
# 청소년 데이터(888) 제거 필요, 무응답(999) 처리 필요

#BS1_1, BS3_1, BS3_2, BS12_37, BS12_47, BS12_47_1 제거
#smoke_ox, smoke_ox2, smoke_ox3 제거
#smoke_sum 변수추가
#==================================
#간접흡연 노출 정도 smoke_second
sum(is.na(data$BS9_2))
sum(is.na(data$BS13))

## 가정내 간접흡연 노출 여부
table(data$BS9_2)
#1. 노출됨 | 2.노출x |  3. 가족 중 본인만 흡연자, 타인에 의해 간접흡연 노출 경험x
#8. 비해당(소아청소년) | 9. 무응답
## 본인이 흡연자면, 직접/간접흡연 다 한다고 봐도 무방할 것 같아 3번 변수는 노출됨으로
봄

##공공장소실내 간접흡연 노출 여부
table(data$BS13)
#1. 노출됨 | 2.노출x  | 8. 비해당(소아청소년) | 9. 무응답

# 간접흡연 노출됨(1), 노출되지 않음(0)으로 위 2가지 변수 새롭게 코딩함
data$smoke_second1 = c()
for (i in 1:dim(data)[1]){
  if(data$BS9_2[i] == 2) {data$smoke_second1[i] <- 0} #간접흡연 노출x
  else if (data$BS9_2[i] == 1 | data$BS9_2[i] == 3) {data$smoke_second1[i] <- 1} #노출o
  else if (data$BS9_2[i] == 8) {data$smoke_second1[i] <- 8} #청소년
  else {data$smoke_second1[i] <- 9} #무응답
}

table(data$smoke_second1)

data$smoke_second2 = c()
for (i in 1:dim(data)[1]){
```

```
  if(data$BS13[i] == 2) {data$smoke_second2[i] <- 0} #간접흡연 노출x
  else if (data$BS13[i] == 1 ) {data$smoke_second2[i] <- 1} #노출o
  else if (data$BS13[i] == 8) {data$smoke_second2[i] <- 8} #청소년
  else {data$smoke_second2[i] <- 9}  #무응답
}
table(data$smoke_second2)

#간접흡연 노출정도를 알아보기 위해,두 변수를 합침
data$smoke_second = c()
for (i in 1:dim(data)[1]){
  if(data$smoke_second1[i]==8| data$smoke_second2[i]==8) {data$smoke_second[i] <- 8}
  else if (data$smoke_second1[i]==9| data$smoke_second2[i]==9) {data$smoke_second[i] <-
9}
  else {data$smoke_second[i] <- data$smoke_second1[i] + data$smoke_second2[i]}
}
table(data$smoke_second)
#0. 간접흡연 노출x | 1. 적게 노출 | 2. 많이 노출  | 8. 비해당(소아청소년) | 9. 무응답


barplot(table(data$smoke_second),main="간접흡연 노출X(0) vs 적게 노출(1) vs 많이
노출(2) vs 비해당(소아청소년)(8) vs무응답(9)")

#BS8_2, BS9_2, BS13, smoke_second1, smoke_second2 제거
#smoke_second 변수추가
#==================================
#앉은 시간 sit
sum(data$BE8_1==88) ; sum(data$BE8_2==88) #비해당(88) 114
sum(data$BE8_1==99) ; sum(data$BE8_2==99) #무응답(99) 359
sum(data$sit==888) ; sum(data$sit==999) #0
#888.999값 원래 없었으므로, 비해당(888), 무응답(999)로 코딩

data$sit <- c()
for (i in 1:dim(data)[1]){
  if(data$BE8_1[i]==88| data$BE8_2[i]==88) {data$sit[i] <- 888} #청소년 비해당
  else if (data$BE8_1[i]==99| data$BE8_2[i]==99) {data$sit[i] <- 999} # 무응답
  else {data$sit[i] <- data$BE8_1[i]*60 + data$BE8_2[i]}
}

hist(data$sit)
sum(data$sit==888) ; sum(data$sit==999) #114,359

# BE8_1, BE8_2 제거, sit 변수 추가
# 무응답 처리 필요
#==================================
#걷기 일수 BE3_31
table(data$BE3_31)
#1.전혀 걷지x | 2.1일 | 3.2일  | 4.3일 | 5.4일 | 6.5일 | 7.6일 | 8.매일 | 88 비해당 | 99 무응답
# 무응답 처리 필요
```

```
#==================================
#운동 일수 BE5_1
table(data$BE5_1)
#1.전혀 운동x | 2.1일 | 3.2일  | 4.3일 | 5.4일 | 6.5일 이상 | 8.비해당(소아) | 9 무응답
# 무응답 처리 필요
#==================================
#부모 질병 의사진단 fh_sum
# 결측치 없음
sum(is.na(data$HE_HPfh1)) ; sum(is.na(data$HE_HPfh2))
sum(is.na(data$HE_HLfh1)) ; sum(is.na(data$HE_HLfh2))
sum(is.na(data$HE_IHDfh1)) ; sum(is.na(data$HE_IHDfh2))
sum(is.na(data$HE_STRfh1)) ; sum(is.na(data$HE_STRfh2))
sum(is.na(data$HE_DMfh1)) ; sum(is.na(data$HE_DMfh2))

#고혈압 HE_HPfh1
table(data$HE_HPfh1)
table(data$HE_HPfh2)

data$HE_HPfh3=c()
for( i in 1:dim(data)[1]){
  if(data$HE_HPfh1[i]==1 | data$HE_HPfh2[i] ==1){
    data$HE_HPfh3[i]=1}
  else if (data$HE_HPfh1[i]==0 & data$HE_HPfh2[i] ==0){
    data$HE_HPfh3[i]=0}
  else data$HE_HPfh3[i] = 9  #0&모름 | 모름&모름
}
table(data$HE_HPfh3)

#고지혈증 HE_HLfh1
table(data$HE_HLfh1)
table(data$HE_HLfh2)

data$HE_HLfh3=c()
for( i in 1:dim(data)[1]){
  if(data$HE_HLfh1[i]==1 | data$HE_HLfh2[i] ==1){
    data$HE_HLfh3[i]=1}
  else if (data$HE_HLfh1[i]==0 & data$HE_HLfh2[i] ==0){
    data$HE_HLfh3[i]=0}
  else data$HE_HLfh3[i] = 9  #0&모름 | 모름&모름
}
table(data$HE_HLfh3)

sum(data$HE_HLfh3==9 |data$HE_HPfh3==9 )

#허혈성심장질환 HE_IHDfh1
table(data$HE_IHDfh1)
table(data$HE_IHDfh2)
```

```
data$HE_IHDfh3=c()
for( i in 1:dim(data)[1]){
  if(data$HE_IHDfh1[i]==1 | data$HE_IHDfh2[i] ==1){
    data$HE_IHDfh3[i]=1}
  else if (data$HE_IHDfh1[i]==0 & data$HE_IHDfh2[i] ==0){
    data$HE_IHDfh3[i]=0}
  else data$HE_IHDfh3[i] = 9  #0&모름 | 모름&모름
}
table(data$HE_IHDfh3)

#뇌졸중 HE_STRfh1
table(data$HE_STRfh1)
table(data$HE_STRfh2)

data$HE_STRfh3=c()
for( i in 1:dim(data)[1]){
  if(data$HE_STRfh1[i]==1 | data$HE_STRfh2[i] ==1){
    data$HE_STRfh3[i]=1}
  else if (data$HE_STRfh1[i]==0 & data$HE_STRfh2[i] ==0){
    data$HE_STRfh3[i]=0}
  else data$HE_STRfh3[i] = 9  #0&모름 | 모름&모름
}
table(data$HE_STRfh3)

#당뇨병 HE_DMfh1
table(data$HE_DMfh1)
table(data$HE_DMfh2)

data$HE_DMfh3=c()
for( i in 1:dim(data)[1]){
  if(data$HE_DMfh1[i]==1 | data$HE_DMfh2[i] ==1){
    data$HE_DMfh3[i]=1}
  else if (data$HE_DMfh1[i]==0 & data$HE_DMfh2[i] ==0){
    data$HE_DMfh3[i]=0}
  else data$HE_DMfh3[i] = 9  #0&모름 | 모름&모름
}
table(data$HE_DMfh3)

sum(data$HE_DMfh3==9 |data$HE_HPfh3==9 )

#정리
data$fh_sum=c()
data$fh_sum= data$HE_HPfh3 +  data$HE_HLfh3 +  data$HE_IHDfh3 +
  data$HE_STRfh3 +  data$HE_DMfh3

table(data$fh_sum)
#HE_HPfh1~3, HE_HLfh1~3, HE_IHDfh1~3, HE_STRfh1~3, HE_DMfh1~3 제거
#fh_sum 변수 추가
```

```
#==================================
#맥압(수축기-이완기 혈압) dis_bp
sum(is.na(data$HE_sbp + data$HE_dbp))

summary(data$HE_sbp)
head(sort(data$HE_sbp))
head(sort(data$HE_sbp, decreasing = TRUE))

summary(data$HE_dbp)
head(sort(data$HE_dbp))
head(sort(data$HE_dbp, decreasing = TRUE))

data$dis_bp <- data$HE_sbp - data$HE_dbp
summary(data$dis_bp)
hist(data$dis_bp)
hist(sqrt(data$dis_bp))
#HE_sbp, HE_dbp 제거, dis_bp 변수 추가
#==================================
#체질량 지수 HE_BMI
sum(is.na(data$HE_BMI))
summary(data$HE_BMI)

hist(data$HE_BMI)
#==================================
#요산도 HE_Uph
sum(is.na(data$HE_Uph))
summary(data$HE_Uph)

hist(data$HE_Uph)
hist(data$HE_Uph**2)
#==================================
#요비중 HE_Usg
sum(is.na(data$HE_Usg))
summary(data$HE_Usg)

hist(data$HE_Usg)
#==================================
#요당 uriglu
sum(is.na(data$HE_Uglu))
table(data$HE_Uglu)
#0.음성 | 1.미량+-  | 2. 양성+  | 3. 양성++ | 4. 양성+++ | 5. 양성++++

data$uriglu <- c()
for (i in 1:dim(data)[1]) {
  if (data$HE_Uglu[i] == 0) {data$uriglu[i] <- 0} # 음성
  else if (data$HE_Uglu[i] == 1) {data$uriglu[i] <- 1} # 미량
  else if (data$HE_Uglu[i] == 2 | data$HE_Uglu[i] == 3 | data$HE_Uglu[i] == 4) {data$uriglu[i]
<- 2} #양성 +,++,+++
```

```
  else {data$uriglu[i] <- 3} #양성++++
}

table(data$uriglu)
#0.음성 | 1.미량+-  | 2. 양성+,++,+++  | 3. 양성++++, 요인 4가지로 나눔


barplot(table(data$uriglu),main="간접흡연 노출X(0) vs 적게 노출(1) vs 많이 노출(2) vs 비해당(소아청소년)(8) vs무응답(9)")

#HE_Uglu 제거
#uriglu 변수 추가
#==================================
#요케톤 HE_Uket
sum(is.na(data$HE_Uket))
table(data$HE_Uket)
#0.음성 | 2. 양성+  | 3. 양성++ | 4. 양성+++

data$HE_Uket2=factor(data$HE_Uket, levels=c(0,2,3,4), labels=c(0,1,2,3))
table(data$HE_Uket2)
#0,2,3,4라서 0,1,2,3으로 순서 바꿔줌


barplot(table(data$HE_Uket2),main="음성(0) vs 양성+(1) vs 양성++(2) vs 양성+++(3)")

#HE_Uket 제거
#HE_Uket2 변수 추가
#==================================
#요잠혈 HE_Ubld
sum(is.na(data$HE_Ubld))
table(data$HE_Ubld)
#0.음성 | 1.미량+-  | 2. 양성+  | 3. 양성++ | 4. 양성+++


barplot(table(data$HE_Ubld),main="음성(0) vs 미량+-(1) vs 양성++(2) vs 양성+++(3)")
#==================================
#식사 빈도 foodfreq
#1.주5~7회 | 2.주3~4회 | 3.주1~2회  | 4. 거의 안 함(주0회) | 5. 무응답

table(data$L_BR_FQ)
table(data$L_LN_FQ)
table(data$L_DN_FQ)
# 식사빈도 데이터 내 모름(5) 빈도 없음

# 식사빈도(아침+점심+저녁) 변수 형성 후, 역코딩
data$foodintake <-  5*3 - (data$L_BR_FQ + data$L_LN_FQ + data$L_DN_FQ)
table(data$foodintake)
## 숫자가 클수록 밥을 하루에 3번 잘 챙겨먹음
```

```
barplot(table(data$foodintake),main="6~9번(0) vs 10~11번(1) vs 12번(2)")
#밥 잘 안 먹는 사람(6~8) 너무 적어, | 0.6~9번 | 1. 10~11번 | 2. 12번 | 3가지 범주로 나눔


data$foodfreq <- c()
for (i in 1:dim(data)[1]) {
  if (data$foodintake[i] <= 9) {data$foodfreq[i] <- 0}
  else if (data$foodintake[i]==10 | data$foodintake[i]==11) {data$foodfreq[i] <- 1}
  else if (data$foodintake[i]==12) {data$foodfreq[i] <- 2}
}

table(data$foodintake)
table(data$foodfreq)
#0. 아침/점심/저녁 잘 챙겨먹지 않음| 1. 식사 적당히 챙겨먹음 | 2. 식사 잘 챙겨먹음


barplot(table(data$foodfreq),main="아침/점심/저녁 잘 챙겨먹지 않음(0) vs 식사 적당히
챙겨먹음(1) vs식사 잘 챙겨먹음(2)")

mosaicplot(data$L_BR_FQ ~ data$foodfreq, main="아침식사과 총식사량")
mosaicplot(data$L_LN_FQ ~ data$foodfreq, main="점심식사와 총식사량")
mosaicplot(data$L_DN_FQ ~ data$foodfreq, main="저녁식사와 총식사량")

#L_BR_FQ,  L_LN_FQ, L_DN_FQ, foodintake 제거
#foodfreq 변수 추가
#==================================
#외식 횟수 eatout
#무응답(9) 없음
sum(is.na(data$L_OUT_FQ))
table(data$L_OUT_FQ)

#외식 횟수 역코딩
data$eatout <- 8 - data$L_OUT_FQ
table(data$eatout)
#1.거의 외식x| 2.월 1~3회 |3.주 1~2회 |4.주 3~4회 |5.주 5~6회 |6.하루 1회 |7.하루 2회 이상
# 숫자가 커질수록 더 많이 외식하는 위와 같은 형태로 역코딩


barplot(table(data$eatout),main="거의 외식x(1) vs월 1~3회(2) vs주 1~2회(3) vs주 3~4회(4)
vs 주 5~6회(5) vs하루 1회(6) vs하루 2회 이상(7)")

#L_OUT_FQ 제거
#eatout 변수 추가
#==================================
#에너지 섭취량 N_EN
sum(is.na(data$N_EN))
```

```
summary(data$N_EN)
head(sort(data$N_EN, decreasing=TRUE))

hist(data$N_EN)
hist(sqrt(data$N_EN))

data$N_EN <- sqrt(data$N_EN)
hist(data$N_EN)
#==================================
#지방 섭취량 fateat
data$fateat <- data$N_FAT + data$N_SFA + data$N_MUFA + data$N_PUFA + data$N_N3
+ data$N_N6 + data$N_CHOL*0.001

sum(is.na(data$fateat))
summary(data$fateat)
head(sort(data$fateat, decreasing=TRUE))

hist(data$fateat)
hist(sqrt(data$fateat))

data$fateat <- sqrt(data$fateat)
hist(data$fateat)


#N_FAT, N_SFA, N_MUFA, N_PUFA, N_N3, N_N6, N_CHOL 제거
#fateat 변수 추가
#==================================
#당 섭취량 N_SUGAR
sum(is.na(data$N_SUGAR))
summary(data$N_SUGAR)
head(sort(data$N_SUGAR, decreasing=TRUE))

hist(data$N_SUGAR)
hist(sqrt(data$N_SUGAR))

data$N_SUGAR <- sqrt(data$N_SUGAR)
hist(data$N_SUGAR)
#==================================
#나트륨 섭취량 N_NA
sum(is.na(data$N_NA))
summary(data$N_NA)
head(sort(data$N_NA, decreasing=TRUE))

hist(data$N_NA)
hist(sqrt(data$N_NA))

data$N_NA <- sqrt(data$N_NA)
hist(data$N_NA)
```

```
 #==================================
```

0.4.1 비해당(소아청소년) 데이터 제거

```
#install.packages("dplyr")
library(dplyr)
dim(data)

#필터 사용자 함수
rm8=c("BP1","BE5_1","DI1_dg","DI2_dg","DE1_dg","DJ4_dg",
      "DL1_dg","DJ8_dg","DJ6_dg",
      "DH4_dg","DN1_dg","BH1", "smoke_second", "weight.ch")
rm88=c("BE3_31")
rm888 = c("smoke_sum", "sit", "sleep")

fil8 <- function(df, variables){
  new=df
  for (variable in variables){
    new <- new%>% filter(new[[variable]] != 8)

  }
  return(new)
}

fil88 <- function(df, variables){
  new=df
  for (variable in variables){
    new <- new%>% filter(new[[variable]] != 88)

  }
  return(new)
}

fil888 <- function(df, variables){
  new=df
  for (variable in variables){
    new <- new%>% filter(new[[variable]] != 888)

  }
  return(new)
}

#비해당 데이터 제거
data <- fil8(data, rm8)
data <- fil88(data, rm88)
data <- fil888(data, rm888)
dim(data) #4638 103
# 전체 데이터 5020개에서 비해당 392개 데이터 제거되어 총 4638개 데이터
```

392개는 반응변수 청소년, 소아에 공통적으로 관측된 값이었음

#0.4.2 범주형 무응답 제거

```
d1<- c("npins",
"sex3","BH1","DJ4_dg","DL1_dg","DJ8_dg","DJ6_dg","DH4_dg","DN1_dg","BE5_1","stress",
"smoke_second","fh_sum")
d2<-c("marri_2","BE3_31")

fil9 <- function(df, variables){
  new=df
  for (variable in variables){
    new <- new%>% filter(new[[variable]] != 9)

  }
  return(new)
}

fil99 <- function(df, variables){
  new=df
  for (variable in variables){
    new <- new%>% filter(new[[variable]] != 99)

  }
  return(new)
}

data <- fil9(data, d1)
data <- fil99(data, d2)
dim(data) # 4296 103

# 추가로, 부모님 의사진단 여부 무응답 수 제거(10~45)
data <- data %>% filter(data$fh_sum < 6)
dim(data) # 3965 103

#=========================================
```

#0.4.3 연속형 무응답(mice)

```
# 범주/순서 무응답 행 제거 후 연속형 MICE 방법으로 무응답 처리
# 근로시간 EC_wht_23 999 | 체중변화 weight.ch -9
# 수면시간 sleep 999 | 앉은 시간 sit 999
table(data$weight.ch)

# mice
#install.packages("mice")
library(mice)
```

```
# 변수를 먼저 결측치로 바꿈
data$EC_wht_23[data$EC_wht_23==999] <- NA
sum(is.na(data$EC_wht_23))

data$weight.ch[data$weight.ch==-9] <- NA
sum(is.na(data$weight.ch))

data$sleep[data$sleep ==999] <- NA
sum(is.na(data$sleep))

data$sit[data$sit ==999] <- NA
sum(is.na(data$sit))
# 무응답 개수 EC_wht_23 4개, weight.ch 1개, sleep 2개로 적어, 해당 변수 행은 제거

fil <- function(df, variables){
  new=df
  for (variable in variables){
    new <- new %>% filter(!is.na(new[[variable]]))

  }
  return(new)
}
k=c("EC_wht_23","weight.ch","sleep")
dim(data) #3965 103
data=fil(data,k)
dim(data) #3958  103

sum(is.na(data$EC_wht_23)) ; sum(is.na(data$weight.ch)) ; sum(is.na(data$sleep))
sum(is.na(data$sit))

# sit 변수 mice 적용
mice_data <- mice(data, seed=2024, meth="pmm", m=10)

completed_data <- complete(mice_data)
print(completed_data)

# 무응답 확인
sum(completed_data$EC_wht_23 == 999) ; sum(is.na(completed_data$EC_wht_23))
sum(completed_data$weight.ch == -9) ; sum(is.na(completed_data$weight.ch))
sum(completed_data$sleep == 999) ; sum(is.na(completed_data$sleep))
sum(completed_data$sit == 999) ; sum(is.na(completed_data$sit))

# 빈도분석
summary(completed_data$EC_wht_23)
summary(completed_data$weight.ch)
summary(completed_data$sleep)
summary(completed_data$sit)
```

#0.4.4 필요없는 변수 제거

```
rm.var=c("BO1_1","BO1_2","BO1_3",
        "BD1", "BD1_11","BD1_11.new","BD2_1","BD2_14", "BD2_1.new",
        "region","sex","LW_pr","BP16_1","BP16_2",
        "BP1","BS1_1", "BS3_1", "BS3_2", "BS12_37", "BS12_47", "BS12_47_1",
        "smoke_ox", "smoke_ox2", "smoke_ox3",
        "BS8_2","BS9_2","BS13","smoke_second1","smoke_second2",
        "BE8_1", "BE8_2",
        "HE_HPfh1", "HE_HPfh2", "HE_HPfh3", "HE_HLfh1", "HE_HLfh2", "HE_HLfh3",
        "HE_IHDfh1", "HE_IHDfh2", "HE_IHDfh3", "HE_STRfh1", "HE_STRfh2", "HE_STRfh3",
        "HE_DMfh1", "HE_DMfh2", "HE_DMfh3",
        "HE_sbp", "HE_dbp",
        "HE_Uglu", "HE_Uket",
        "N_WAT_C",
        "L_BR_FQ","L_LN_FQ","L_DN_FQ","foodintake",
        "L_OUT_FQ",
        "N_FAT", "N_SFA", "N_MUFA", "N_PUFA", "N_N3", "N_N6", "N_CHOL")

length(rm.var) #63
dim(data) #3958 103

library(tidyverse)
completed_data= select(completed_data, - rm.var)
completed_data
dim(completed_data) #3958 40

head(completed_data)

# write.csv(completed_data , file = "C:/Users/DS/Downloads/completed_data.csv",
row.names = F)
```

#1.0 결측치 없는 완전한 데이터 completed_data

```
setwd("C:/Users/DS/Downloads")
completed_data <- read.csv("completed_data.csv")
f.name=c("city","sex3","marri_2","npins","BH1")
completed_data[, f.name]

#to.factors <- function(df, variables){
# for (variable in variables){  df[[variable]] <- as.factor(df[[variable]])   }
#  return(df)  }

completed_data=to.factors(completed_data, f.name)
str(completed_data)
dim(completed_data) #3958  40
```

#1.1. 단계적 선택방법(로지스틱, 포아송)

```
## 로지스틱
model <- glm(DI1_dg ~., family = binomial(link=logit), data = completed_data)
result <- step(model, direction="both")

model2 <- glm(DI2_dg ~., family = binomial(link=logit), data = completed_data)
result2 <- step(model2, direction="both")

model3 <- glm(DE1_dg ~., family = binomial(link=logit), data = completed_data)
result3 <- step(model3, direction="both")

model4 <- glm(DJ4_dg ~., family = binomial(link=logit), data = completed_data)
result4 <- step(model4, direction="both")

model5 <- glm(DL1_dg ~., family = binomial(link=logit), data = completed_data)
result5 <- step(model5, direction="both")

model6 <- glm(DJ8_dg ~., family = binomial(link=logit), data = completed_data)
result6 <- step(model6, direction="both")

model7 <- glm(DJ6_dg ~., family = binomial(link=logit), data = completed_data)
result7 <- step(model7, direction="both")

model8 <- glm(DH4_dg~., family = binomial(link=logit), data = completed_data)
result8 <- step(model8, direction="both")

model9 <- glm(DN1_dg~., family = binomial(link=logit), data = completed_data)
result9 <- step(model9, direction="both")

#단계적 선택방법 결과, 유의한 변수들이 다른 반응변수로 사용할 질병 변수를 대부분
포함하고 있었음. 질병간의 관계가 높다고 판단하여, 각 질병을 반응변수로 보지 않고,
질병들을 하나로 합쳐 포아송 형태로 반응변수를 선택하기로 결정
#==================================
##포아송
#고혈압~콩팥병 진단여부 빈도
dg_sum=completed_data$DI1_dg + completed_data$DI2_dg +
completed_data$DE1_dg + completed_data$DJ4_dg  +
 completed_data$DL1_dg  +  completed_data$DJ6_dg  +
 completed_data$DJ8_dg  +  completed_data$DH4_dg  +
 completed_data$DN1_dg

table(dg_sum) #0~6까지 존재
dg_sum[dg_sum>=4]=4
table(dg_sum) #0~4까지 존재 #빈도가 적어서 합침

sum_data=data.frame(completed_data[,c(1:4,14:40)],dg_sum)
head(sum_data)
```

```
table(sum_data$dg_sum)

model <- glm(dg_sum ~., family = poisson, data = sum_data)
result <- step(model, direction="both")
#유의한 변수: age, BH1, EC_wht_23, BE3_31, HE_BMI, N_SUGAR, sex3, weight.ch, BD.total,
sleep, stress, sit, fh_sum, dis_bp, uriglu


# write.csv(sum_data , file = "C:/Users/DS/Downloads/sum_data.csv", row.names = F)
#==================================

#1.2. 반응변수: 질병진단 개수에 대한 연속형변수 데이터
setwd("C:/Users/DS/Downloads")
sum_data=read.csv("sum_data.csv", header = T)
f.name=c("city","sex3","marri_2","npins","BH1")

to.factors <- function(df, variables){
  for (variable in variables){  df[[variable]] <- as.factor(df[[variable]])   }
  return(df)  }

sum_data=to.factors(sum_data, f.name)
str(sum_data)

dim(sum_data) #3958 32

 #==================================

#1.3 설명변수 간 관계

#1.3.1 연속형/순서형 설명변수 상관관계
#상관계수 절대값 0.4 이상인 변수들 조합
seq <- read.csv("seq_name.csv")
seq_data <- sum_data[, seq$name]

sum(cor(seq_data)!=1 & cor(seq_data)>=0.4) #14
sum(cor(seq_data)!=1 & cor(seq_data)<=-0.4) #2

sum(cor(seq_data, method="spearman")!=1 & cor(seq_data, method="spearman")>=0.4)
#14
sum(cor(seq_data, method="spearman")!=1 & cor(seq_data, , method="spearman")<=-0.4)
#2
#연속형 변수-순서형 변수에는 0.4 넘는 거 존재x
#오로지 연속형 변수-연속형 변수에만 존재

cor(seq_data)[cor(seq_data)!=1 & cor(seq_data)>=0.4]
cor(seq_data)[cor(seq_data)!=1 & cor(seq_data)<=-0.4]
```

```
cor(seq_data)!=1 & cor(seq_data)>=0.4
cor(seq_data)!=1 & cor(seq_data)<=-0.4

#age-dis_bp(0.4309), age-foodfreq(0.4659), EC_wht_23-eatout(0.4184)
#N_EN-fateat(0.7573), N_EN-N_SUGAR(0.5471), N_EN-N_NA(0.6667),
fateat-N_NA(0.4881)
#age-eatout(-0.4829)
 #==================================
#상관분석 그림
 # 피어슨 상관분석(연속-연속)
library(corrplot)
contin_data <- select(sum_data, c("eatout", "ainc", "EC_wht_23", "weight.ch",
                            "BD.total", "sleep", "smoke_sum", "sit",
                            "dis_bp", "HE_BMI", "HE_Uph", "HE_Usg",
                            "N_SUGAR", "age", "foodfreq", "N_EN", "fateat", "N_NA"))

corrplot(cor(contin_data), method="shade", shade.col=NA, tl.col="black", tl.srt=35,
col=col(200), addCoef.col="black", order="AOE") #사각형

corrplot(cor(contin_data), type="lower",
        method="shade", shade.col=NA, tl.col="black", tl.srt=8, col=col(200),
addCoef.col="black", order="AOE") #삼각형

 #==================================

# 스피어만 상관분석(순서-순서)
library(corrplot)
ord_data <- select(sum_data, c("stress", "smoke_second",
                            "BE3_31", "BE5_1", "fh_sum", "uriglu",
                            "HE_Uket2","HE_Ubld"))
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))

corrplot(cor(ord_data, method="spearman"), method="shade", shade.col=NA, tl.col="black",
tl.srt=35, col=col(200), addCoef.col="black", order="AOE") #사각형
corrplot(cor(ord_data, method="spearman"), type = "lower",
        method="shade", shade.col=NA, tl.col="black", tl.srt=35, col=col(200),
addCoef.col="black", order="AOE") #삼각형
 #==================================
# 스피어만 상관분석(순서-연속)
ord_data <- select(sum_data, c("stress", "smoke_second",
                            "BE3_31", "BE5_1", "fh_sum", "uriglu",
                            "HE_Uket2","HE_Ubld"))

ord_contin_data <- select(sum_data, c("stress",
"eatout", "ainc", "EC_wht_23", "weight.ch", "BD.total", "sleep", "smoke_sum", "sit", "dis_bp",
"HE_BMI", "HE_Uph", "HE_Usg", "N_SUGAR", "age", "foodfreq", "N_EN", "fateat", "N_NA"))
```

```
corrplot(cor(ord_contin_data, method="spearman"), method="shade", shade.col=NA,
tl.col="black", tl.srt=8, col=col(200), addCoef.col="black", order="AOE")
## 순서형 변수 하나씩 넣어서 상관계수 확인
 #====================================

#1.3.2 범주형 설명변수 간 관계
#결혼과 보험가입여부(marri_2,npins)
library(gmodels)
chisq.test(sum_data$marri_2, sum_data$npins)
CrossTable(sum_data$marri_2, sum_data$npins)
#결혼과 보험가입여부 < 2.2e-16로 관계 있음
 #====================================
#결혼과 건강검진여부(marri_2,BH1)
chisq.test(sum_data$marri_2, sum_data$BH1)
CrossTable(sum_data$marri_2, sum_data$BH1)
#결혼과 건강검진여부 < 2.2e-16로 관계 있음
 #====================================
#결혼과 임신여성/임신x여성/남성(marri_2,sex3)
chisq.test(sum_data$marri_2, sum_data$sex3)
CrossTable(sum_data$marri_2, sum_data$sex3)
#결혼과 임신여성/임신x여성/남성 변수는 < 2.2e-16로 관계 있음
 #====================================
#보험가입여부와 수도권 지역(npins, city)
chisq.test(sum_data$npins, sum_data$city)
CrossTable(sum_data$npins, sum_data$city)
#보험가입과 도시 유의확률  5.58e-15로 유의
 #====================================
#보험가입여부와 임신/여성/남성 (npins, sex3)
chisq.test(sum_data$npins, sum_data$sex3)
CrossTable(sum_data$npins, sum_data$sex3)
#보험가입과 임신/여성/남성 유의확률  0.0001736로 유의
 #====================================
#건강검진여부와 수도권(BH1, city)
chisq.test(sum_data$BH1, sum_data$city)
CrossTable(sum_data$BH1, sum_data$city)
#p값=1로 유의x, 건강검진여부와 수도권은 관련 없음
 #====================================
#건강검진여부와 임신/여성/남성(BH1, sex3)
chisq.test(sum_data$BH1, sum_data$sex3)
CrossTable(sum_data$BH1, sum_data$sex3)
#p값=p-value = 1.875e-15로, 건강검진여부와 임신/여성/남성 관련있음
 #====================================
#도시와 임신/여성/남성(city, sex3)
chisq.test(sum_data$city, sum_data$sex3)
CrossTable(sum_data$city, sum_data$sex3)
#p값 4.719e-08로 유의, 도시와 임신/여성/남성 관련있음
 #====================================
```

```
#건강검진여부와 보험가입여부(BH1, npins)
chisq.test(sum_data$BH1, sum_data$npins)
CrossTable(sum_data$BH1, sum_data$npins)
# p값 1.899e-06로 유의, 건강검진여부와 보험가입여부 관련있음
#==================================
#도시와 결혼상태(city, marri_2)
chisq.test(sum_data$city, sum_data$marri_2)
CrossTable(sum_data$city, sum_data$marri_2)
#p값 1.379e-15로 유의, 도시와 결혼상태 관련있음
#==================================
```

#1.3.3 주성분분석

```
#PCA
n.name=read.csv("seq_name.csv", header=T)
n.name$name[-3] #반응변수 제외

n.data=sum_data[,n.name$name[-3] ] #반응변수 제외. 연속형, 순서형 데이터
n.data

pca= prcomp(n.data , center=T , scale. = T)  #표준화된 데이터 사용
pca
summary(pca)
#Cumulative Proportion PC17까지는 가야지 0.8이상이됨

screeplot(pca, type="lines", pch=19, main="Scree plot")
#급격하게 완만해지는 지점으로 주성분 선택
#10까지만 결과가 보여짐. 4부터 완만해지기 시작

#고유값 확인
pca$sdev^2
#고유값이 0.7 이상이어야 한다
pca$sdev^2 >= 0.7
#17까지는 사용가능

#각 주성분이 나타내는 것
#PC1
for( i in 1:17){
print( pca$rotation[order(abs(pca$rotation[,i])),i] [c(26,25)] )
}

#주성분 점수를 이용해서 변수 생성
k=17
pca$rotation[,c(1:k)]
pc=as.matrix(n.data) %*% pca$rotation[,c(1:k)]

str(pc) #행렬곱의 결과이기 때문에 행렬형태
pc=data.frame(pc) #데이터프레임으로 변환
```

```
str(pc)
#주성분 분석 사용해서, 변수 줄이는 게 어려울 것 같아 사용하지 않기로 함
 #==================================
```

#1.3.4 연속형/순서형 요인분석

```
#순서/연속형 데이터 표준화
str(seq_data[, -3]) #반응변수 제외
seq_data2 <- seq_data[, -3]
seq_data2 <- data.frame(scale(seq_data2))
summary(seq_data2)
 #==================================
#적절한 변수 탐색(고유값, screeplot)
seq_en <- eigen(cor(seq_data2))
seq_en$values # 고유값 1 이상 10개

plot(seq_en$values, typ="o", main="scree plot")
 #==================================
#요인분석(고유값 변수 10개 기준)
seq_result <- factanal(seq_data2, factors=10, rotation="varimax")
seq_result
print(seq_result, cut=0.5)
#(N_EN, fateat, N_NA), (age, foodfreq),  (BD.total, smoke_sum) 서로 묶임
#==================================
#요인분석(p값 유의 13개 기준)
seq_result2 <- factanal(seq_data2, factors=13, rotation="varimax")
seq_result2
print(seq_result2, cut=0.5)
#(N_EN, fateat, N_NA), (age, foodfreq) 서로 묶임
#==================================
#공통적으로 나온 변수 (N_EN, fateat, N_NA), (age, foodfreq)만 가지고 다시 요인분석
foodagedata <- seq_data2[, c("N_EN","fateat","N_NA", "age", "foodfreq")]
seq_en <- eigen(cor(foodagedata))
seq_en$values

result <- factanal(foodagedata, factors=2)
result
#p값>0.05이므로, 유의수준 0.05에서 모형 유의한 결과 나왔으며, Loadings 값 0.5 이상으로 판단했을 때, 변수도 (N_EN, fateat, N_NA)와 (age, foodfreq)로 잘 나누어졌음.
#변수 2개로 나누었을 때 누적 분산 비율이 64.2%로 데이터를 잘 설명함, 변수 합쳐서 사용
#==================================
```

#1.3.5 sum_data2(요인분석 후 합친 변수 있는 버전) 생성

```
sum_data2 = sum_data
dim(sum_data2) ; dim(sum_data) #3958 32

sum_data2$sumption <- (sum_data$N_EN + sum_data$fateat + sum_data$N_NA)/3
```

```
sum_data2$agefood <- scale(sum_data$age) + scale(sum_data$foodfreq)

dim(sum_data2) #3958 34

str(sum_data2)
# 새로운 변수 sumption, agefood 존재, 새로운 변수 만들 때 사용했던 변수들 제거

rm.var = c("N_EN", "fateat", "N_NA",
           "age", "foodfreq")

library(tidyverse)
sum_data2 <- select(sum_data2,-rm.var)
dim(sum_data) #3958 32
dim(sum_data2) #3958 29
#==================================
#요인분석 후 연속형 상관관계 해석
contin_data2 <- select(sum_data2, c("eatout", "ainc", "EC_wht_23", "weight.ch",
                                    "BD.total", "sleep", "smoke_sum", "sit",
                                    "dis_bp", "HE_BMI", "HE_Uph", "HE_Usg",
                                    "N_SUGAR", "agefood", "sumption"))

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(cor(contin_data2), method="shade", shade.col=NA, tl.col="black", tl.srt=35,
col=col(200), addCoef.col="black", order="AOE") #사각형

corrplot(cor(contin_data2), type="lower",
         method="shade", shade.col=NA, tl.col="black", tl.srt=8, col=col(200),
addCoef.col="black", order="AOE") #삼각형

#==================================
#age와 foodfreq 변수 관계 확인

# boxplot
boxplot(age ~ foodfreq, data=sum_data)

# 연령대 구간 설정
sum_data3 <- sum_data
age_intervals <- cut(sum_data3$age, breaks = c(0, 30, 40, 50, 60, Inf),
                     labels = c("30세 이하", "31~40", "41~50", "51~60세", "60세 이상"))

# 데이터에 연령대 정보 추가
sum_data3$age_group <- factor(age_intervals, levels = c("30세 이하", "31~40", "41~50",
"51~60세", "60세 이상"))
table(sum_data3$age_group)



# ggplot을 이용한 연령대별 boxplot 그리기
```

```
ggplot(sum_data3, aes(x = age_group, y = foodfreq)) +
  geom_boxplot(color = "black", fill = "lightblue") +  # 상자 그림 설정
  labs(x = "Age Group", y = "Food Frequency") +  # 축 이름 설정
  theme_minimal()  # 최소한의 배경 설정
#==================================
```

#1.3.6 train, test 분할

```
test <- sample(1:dim(sum_data)[1], size = dim(sum_data)[1]*0.3)
test_set1 <- sum_data[test, ]
train_set1 <- sum_data[-test, ]

head(test)

table(test_set1$dg_sum)
table(train_set1$dg_sum)
table(sum_data$dg_sum)

#write.csv(test , file = "C:/Users/Kwon/Desktop/Downloads/test.csv", row.names = F)
#random으로 돌리지 말고, train, test set 분할 되어있는 index 엑셀 파일 사용

# train, test set 분할 되어있는 엑셀 사용
test <- read.csv("test.csv")
test
#==================================
```

#1.3.7 요인분석 전 Lasso 분석

```
x <- model.matrix(dg_sum ~ ., sum_data)[,-1]
y <- sum_data$dg_sum

head(x) #요인분석 전 데이터 맞음

library(glmnet)
grid <- 10^seq(10, -2, length = 100)

train <- -test$x
y.test <- y[test$x]
y.test

lasso.mod <- glmnet(x[train, ], y[train], alpha = 1,  lambda = grid)
plot(lasso.mod, label=TRUE)
names.vector = lasso.mod$beta[,length(lasso.mod$lambda)]
legend("topleft", legend = names(names.vector), col=1:length(names.vector), lty=1, cex=0.4)



set.seed(1)
cv.out <- cv.glmnet(x[train, ], y[train], alpha = 1)
```

```
plot(cv.out)

bestlam <- cv.out$lambda.min
bestlam #0.008047023

lasso.pred <- predict(lasso.mod, s = bestlam, newx = x[test$x, ])
mean((lasso.pred - y.test)^2) #0.9749766

out <- glmnet(x, y, alpha = 1, lambda = grid)
lasso.coef <- predict(out, type = "coefficients",  s = bestlam)[,]
lasso.coef
lasso.coef[lasso.coef != 0]
#age, ainc, marri_23, marri_24, BH11, EC_wht_23, BE3_31, HE_BMI, HE_Uph, N_SUGAR, N_NA
#city1, sex31, sex32, weight.ch, BD.total, sleep, stress, sit
#fh_sum, dis_bp, uriglu,foodfreq (sumption, agefood 제거) 유의한 변수들
#요인분석 후 lasso와 비교하면, agefood(age, foodfreq)로 들어오고, sumption(N_EN, N_NA, fateat)인데 N_NA만 들어옴
# 랜덤이나, 만들어진 test index 엑셀 사용하나 변수 선택 결과 동일, test index 사용
#==================================

#1.4 반응변수와 설명변수 관계(수정완료)

#1.4.1 범주형 설명변수 - 순서형 반응변수
#이분형 설명변수 - 순서형 반응변수 : T검정
# levene 등분산 검정
library(car)
leveneTest(sum_data$dg_sum, sum_data$city, center=mean, data=sum_data)
leveneTest(sum_data$dg_sum, sum_data$npins, center=mean, data=sum_data)
leveneTest(sum_data$dg_sum, sum_data$BH1, center=mean, data=sum_data)
#city, BH1 등분산성 가정 만족, npins 등분산성 가정 만족x

# 독립 t-test 진행
t.test(sum_data$dg_sum~sum_data$city, data=sum_data, var.equal=T)
t.test(sum_data$dg_sum~sum_data$npins, data=sum_data, var.equal=F)
t.test(sum_data$dg_sum~sum_data$BH1, data=sum_data, var.equal=T)
#반응변수와 city 변수는 관계O, 0(비수도권)이 1(수도권)보다 질병 걸릴 확률 높음
#반응변수와 npins 변수 관계O, 2(보험 가입x)가 1(보험 가입O)보다 질병 걸릴 확률이 높음
#반응변수와 BH1 변수 관계O, 1(건강검진함)이 0(건강검진안함)보다 질병 걸릴 확률 높음
#==================================
#범주 3개 이상 설명변수 - 순서형 반응변수 : 일원배치 분산분석
plot(dg_sum~sex3, data=sum_data) #0,1 범주 비슷, 2가 질병 여부 낮음
plot(dg_sum~marri_2, data=sum_data) #1,2,3 범주 비슷, 4가 질병 여부 낮음

sex3_aov <- aov(dg_sum~sex3, sum_data)
summary(sex3_aov)
#sex3와 반응변수 관련O
```

```
marri_2_aov <- aov(dg_sum~marri_2, sum_data)
summary(marri_2_aov)
#marri_2와 반응변수 관련O

plot(sex3_aov) #정규성 만족, 잔차의 독립성, 등분산성 괜찮아보임
plot(marri_2_aov) # 정규성 만족, sex3보다 독립성, 등분산성 불만족

bartlett.test(dg_sum~sex3, data=sum_data) #등분산성 만족
bartlett.test(dg_sum~marri_2, data=sum_data) #등분산성 불만족

# 사후분석
pairwise.t.test(sum_data$dg_sum, sum_data$sex3, p.adjust="bonferroni", pool.sd=TRUE)
#sex3은 0(남자)와 1(여자와 임신o)가 묶이고 2(여자와 임신x)는 0,1 변수와 다른 값

pairwise.t.test(sum_data$dg_sum, sum_data$marri_2, p.adjust="bonferroni",
pool.sd=FALSE)
# 범주 (1,2)묶이고,3,4가 질병 정도가 다르다고 봐야할 것 같음
#==================================
```

#1.4.2 반응변수와 연속/순서형 상관분석

```
ord_data <- select(sum_data, c("dg_sum", "stress", "smoke_second",
                               "BE3_31", "BE5_1", "fh_sum", "uriglu",
                               "HE_Uket2","HE_Ubld", "eatout"))

str(ord_data)

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(cor(ord_data, method="spearman"), method="shade", shade.col=NA, tl.col="black",
tl.srt=35, col=col(200), addCoef.col="black", order="AOE") #사진첨부
#반응변수와 각 순서형 설명변수의 상관계수는 stress(-0.01), smoke_second(0), BE3_31(0.01),
BE5_1(-0.02), fh_sum(0.06), uriglu(0.18), HE_Uket2(-0.04), HE_Ubld(0.01), eatout(-0.14)로 낮은 편임

cor.test(sum_data$dg_sum, sum_data$age, method="spearman")
cor.test(sum_data$dg_sum, sum_data$foodfreq, method="spearman")
cor.test(sum_data$dg_sum, sum_data$N_EN, method="spearman")
cor.test(sum_data$dg_sum, sum_data$N_NA, method="spearman")
cor.test(sum_data$dg_sum, sum_data$fateat, method="spearman")
#반응변수와 요인분석에서 합쳐지는 기존 변수의 상관계수는 age(0.2668658), foodfreq(0.130552),
N_EN(-0.04958949), N_NA(-0.02411277), fateat(-0.09272031)임, 다 tie존재해서 정확한 값 계산x

contin_data <- select(sum_data2, c("dg_sum", "ainc", "EC_wht_23", "weight.ch",
                                   "BD.total", "sleep", "smoke_sum", "sit",
                                   "dis_bp", "HE_BMI", "HE_Uph", "HE_Usg",
                                   "N_SUGAR", "agefood", "sumption"))
str(contin_data)
corrplot(cor(contin_data, method="spearman"), method="shade", shade.col=NA,
tl.col="black", tl.srt=35, col=col(200), addCoef.col="black", order="AOE") #사진첨부
```

#반응변수와 각 연속형 설명변수의 상관계수는 ainc(-0.14), Ec_wht_23(-0.12), weight.ch(-0.06), BD.total(-0.12), sleep(-0.05), smoke_sum(-0.03), sit(0.05), dis_bp(0.21), HE_BMI(0.17), HE_Uph(0.02), HE_Usg(-0.05), N_SUGAR(-0.04), agefood(0.27), sumption(-0.04)로 낮은 편임
#전체적으로 반응변수 dg_sum과 연속/순서형 설명변수의 상관계수가 낮은 편임

#==================================

#2. 모형 비교

#2.1 데이터 부르기

```
#sum_data(기존 데이터)
setwd("C:/Users/DS/Downloads")
sum_data=read.csv("sum_data.csv", header = T)
f.name=c("city","sex3","marri_2","npins","BH1")

to.factors <- function(df, variables){
  for (variable in variables){  df[[variable]] <- as.factor(df[[variable]])   }
  return(df)  }

sum_data=to.factors(sum_data, f.name)
str(sum_data)

dim(sum_data) #3958 32

#sum_data2(요인분석 완료 데이터)
sum_data2 = sum_data
dim(sum_data2) ; dim(sum_data) #3958 32

sum_data2$sumption <- (sum_data$N_EN + sum_data$fateat + sum_data$N_NA)/3
sum_data2$agefood <- scale(sum_data$age) + scale(sum_data$foodfreq)

dim(sum_data2) #3958 34

str(sum_data2)
# 새로운 변수 sumption, agefood 존재, 새로운 변수 만들 때 사용했던 변수들 제거

rm.var = c("N_EN", "fateat", "N_NA",
           "age", "foodfreq")

library(tidyverse)
sum_data2 <- select(sum_data2,-rm.var)
dim(sum_data) #3958 32
dim(sum_data2) #3958 29

#==================================
```

#2.2 데이터 train, test 분할

```
test <- read.csv("test.csv")
test

sum_data_test <- sum_data[test$x, ]
sum_data_train <- sum_data[-test$x, ]

sum_data2_test <- sum_data2[test$x, ]
sum_data2_train <- sum_data2[-test$x, ]

table(sum_data_test$dg_sum)
table(sum_data_train$dg_sum)
table(sum_data$dg_sum)

dim(sum_data) #3958, 32
dim(sum_data2) #3958, 29
#==================================
```

#2.3 기존 데이터 포아송 모형(모델1)

```
mean(sum_data$dg_sum) ; var(sum_data$dg_sum)
#1.019454 / 1.145692, 평균과 분산 차이 크지 않으므로 기본 포아송 사용(음이항x)

model1 <- glm(dg_sum ~., family = poisson, data = sum_data_train)
summary(model1)
#AIC=7124.7, Null deviance(3615.6, df=2770), Residual deviance(3144.7, df=2736)

pre1 = predict(model1, sum_data_test, type="response")
mse1=sum((sum_data_test$dg_sum - pre1)**2)/dim(sum_data_test)[1]

pre_round1 <- round(pre1)
mse1_r=sum((sum_data_test$dg_sum - pre_round1)**2)/dim(sum_data_test)[1]

hist(pre1)
hist(pre_round1)
#0~1사이의 많은 값들이, 다 1로 분류되어 오류 증가함
#==================================
```

#2.4 요인분석 포아송 모형(모델2)

```
model2 <- glm(dg_sum ~., family = poisson, data = sum_data2_train)
summary(model2)
#AIC=7161, Null deviance(3615.6, df=2770), Residual deviance(3187.0, df=2739)

pre2 = predict(model2, sum_data2_test, type="response")
mse2=sum((sum_data2_test$dg_sum - pre2)**2)/dim(sum_data2_test)[1]
mse2
#==================================
```

#2.5 요인분석+단계적 포아송 모형(모델3)

```
model3 <- step(model2, direction="both")
#Step:  AIC=7140.41
#dg_sum ~ BH1 + EC_wht_23 + BE3_31 + HE_BMI + HE_Uph + sex3 +
#weight.ch + BD.total + stress + sit + fh_sum + dis_bp + uriglu +
#sumption + agefood

summary(model3)
#AIC=7140.4, Null diviance(3615.6, df=2770), Residual deviance(3196.4, df=2754)

pre3 = predict(model3, sum_data2_test, type="response")
mse3= sum((sum_data2_test$dg_sum - pre3)**2)/dim(sum_data2_test)[1]
mse3
#===================================
```

#2.6 lasso 선택 모형(모델4)

```
lasso_var <- c("age", "ainc", "marri_2", "BH1", "EC_wht_23", "BE3_31",
               "HE_BMI", "HE_Uph", "N_SUGAR", "N_NA", "city", "sex3",
               "weight.ch", "BD.total", "sleep", "stress", "sit",
               "fh_sum", "dis_bp", "uriglu", "foodfreq", "dg_sum")

lasso_train <- select(sum_data_train, lasso_var)
lasso_test <- select(sum_data_test, lasso_var)

model4 <- glm(dg_sum ~., family = poisson, data = lasso_train)
summary(model4)
#AIC=7110.8, Null deviance(3615.6, df=2770), Residual deviance(3150.8, df=2746)

pre4 = predict(model4, lasso_test, type="response")
mse4=sum((lasso_test$dg_sum - pre4)**2)/dim(lasso_test)[1]
mse4
#===================================
```

#3. 모형 비교

```
cbind(mse1, mse2, mse3, mse4)
#===================================
```

#3.1 모형 카이제곱

```
b = summary(model1)
pchisq(b$null.deviance - b$deviance, b$df.null - b$df.residual, lower.tail=F)
#p=2.554398e-78, 회귀계수가 적어도 하나 유의하다.

b = summary(model2)
pchisq(b$null.deviance - b$deviance, b$df.null - b$df.residual, lower.tail=F)
#p=1.70988e-71, 회귀계수가 적어도 하나 유의하다.
```

```
b = summary(model3)
pchisq(b$null.deviance - b$deviance, b$df.null - b$df.residual, lower.tail=F)
#p=3.437512e-79, 회귀계수가 적어도 하나 유의하다.

b = summary(model4)
pchisq(b$null.deviance - b$deviance, b$df.null - b$df.residual, lower.tail=F)
#p=3.347881e-83, 회귀계수가 적어도 하나 유의하다.

#==================================

#3.2. anova모형비교
anova(model3, model2, test="Chisq") #p=0.8542, 모형이 다르지 않다.
anova(model4, model1, test="Chisq") #p=0.8048827, 모형이 다르지 않다.

anova(model3, model4, test="Chisq") #p=2.856e-07, 모형이 다르다.
```