중급보건통계학 과제 2 (고콜레스테롤혈증)

2022년 2학기 중급보건통계학 Take Home Exam

제출일: 2022-12-11

이름: 김연주

학과: 보건정책관리학부

학번: 2021250461

1. 전처리 과정: 대상 변수는 age, sex, HE_chol, HE_wc, BS3_1, BD1, HE_ht, HE_wt, HE_HCHOL.

1-1. 40세 이상으로 데이터셋을 만들어 분석대상을 한정하시오(n=4,488 확인).

- **Script:** *knh20_1=subset(knh20,age>=40)*

- **설명**: import dataset 메뉴를 통해 knh20.csv 파일을 읽었으며, 40세 이상의 관측치만으로 구성된 새로운 데이터셋 "knh20_1"을 생성하였다. N=4,488임도 확인하였다.

- **N=4488 확인**

1-2. 4-5페이지의 표를 참고하여 범주형 변수(BD1)에 대하여 결측값(8,9 등)을 빈도분석으로 확인하시오. 이들 값이 있는 경우에 결측치(NA)로 처리하시오.(변수명은 동일하게 함) 정확하게 처리되었는지 결측값 처리 전후를 비교하시오.

- **Script:**

library(prettyR)

freq(knh20_1$BD1)

knh20_1$BD1=as.factor(ifelse(knh20_1$BD1<8,knh20_1$BD1,NA))

freq(knh20_1$BD1)

- **설명**: 빈도를 확인한 결과 8로 관측된 값은 존재하지 않았고, 9로 응답되어 있는 관측치가 53개 존재하였다. 해당 관측치를 모두 결측값으로 처리하였고 변수명은 변경하지 않았다. 결측값 처리 이후 다시 빈도를 확인하였고, 처리가 올바르게 되었음을 알 수 있었다.

- **결측값 처리 이전 빈도**: Frequencies for knh20_1$BD1

```
        2    1    9   NA
     3641  657   53  137
%    81.1 14.6  1.2  3.1
%!NA 83.7 15.1  1.2
```

- **결측값 처리 이후 빈도:** Frequencies for knh20_1$BD1

```
         2    1   NA
      3641  657  190
%     81.1 14.6  4.2
%!NA 84.7 15.3
```

1-3. BS3_1(현재흡연) 변수의 분포를 확인하고 비흡연(8), 과거흡연(3), 현재흡연(1,2)으로 재분류하고 결측값(9)를 처리하여 smoke(흡연상태, 1=비흡연, 2=과거흡연, 3=현재흡연) 변수를 생성하시오. 또한 BS3_1와 smoke 빈도를 비교하여 재분류가 정확하게 되었는지 확인하시오.

- **Script:**

```
freq(knh20_1$BS3_1)
knh20_1$smoke=ifelse(knh20_1$BS3_1<=2,3,ifelse(knh20_1$BS3_1==3,2,ifelse(knh20_1$BS3_1==8,1,ifelse(knh20_1$BS3_1<=8,knh20_1$BS3_1,NA))))
freq(knh20_1$smoke)
```

- **설명**: knh20_1 데이터셋의 BS3_1 변수의 빈도를 확인하였다. 이후 ifelse 문을 이용하여 smoke 변수를 생성하였고, smoke 변수의 빈도를 다시 확인하여 분류가 올바르게 되었음을 알 수 있었다.

- **BS3_1 변수의 빈도**: Frequencies for knh20_1$BS3_1

```
         8    3    1    2    9   NA
      2570 1064  589   71   57  137
%     57.3 23.7 13.1  1.6  1.3  3.1
%!NA 59.1 24.5 13.5  1.6  1.3
```

- **Smoke 변수의 빈도**: Frequencies for knh20_1$smoke

```
         1    2    3   NA
      2570 1064  660  194
%     57.3 23.7 14.7  4.3
%!NA 59.9 24.8 15.4
```

1-4. 전처리가 완료된 상태에서 age, sex, HE_chol, HE_wc, smoke, BD1, HE_ht, HE_wt, HE_HCHOL 변수를 포함하는 subset을 만들고 변수들의 기술통계를 보고하시오. 연속변수는 n, 평균, SD, 최소, 최대를, 범주형 변수는 빈도와 %를 보고하시오. (단, 결측치가 있는 표본을 별도로 제거하지 않고 포함하여 n=4,488로 유지)

- **Script:**

```
knh20_2=subset(knh20_1,select=c(age,sex,HE_chol,HE_wc,smoke,BD1,HE_ht,HE_wt,HE_HCHOL))

library(psych)
describe(knh20_2$age)
describe(knh20_2$HE_chol)
describe(knh20_2$HE_wc)
describe(knh20_2$HE_ht)
describe(knh20_2$HE_wt)

library(prettyR)
freq(knh20_2$sex)
freq(knh20_2$smoke)
freq(knh20_2$BD1)
freq(knh20_2$HE_HCHOL)
```

- **설명**: age, sex, HE_chol, HE_wc, smoke, BD1, HE_ht, HE_wt, HE_HCHOL 변수를 포함하는 새로운 데이터셋 "knh20_2"를 생성하였다. 이후 연속변수인 age, HE_chol, HE_wc, HE_ht, HE_wt는 describe문을 사용하고, 범주형 변수인 sex, smoke, BD1, HE_HCHOL은 freq문을 사용하여 기술통계를 구하였다. 이때 knh20_2의 관측값 개수는 4,488임도 확인하였다.

- **N=4488 확인**

- **연속형 변수 기술**

	n	평균	SD	최소	최대
age	4488	60.56	11.87	40	80
HE_chol	4240	188.99	40.6	61	418

HE_wc	4330	85.85	9.89	29.8	127.3
HE_ht	4264	161.79	9.2	52	198
HE_wt	4336	63.73	11.96	32.7	125.2

- **범주형 변수 빈도**

```
Sex: Frequencies for knh20_2$sex
        2    1   NA
     2497  1991   0
%    55.6  44.4    0
%!NA 55.6  44.4

Smoke: Frequencies for knh20_2$smoke
        1    2    3   NA
     2570 1064  660  194
%    57.3 23.7 14.7  4.3
%!NA 59.9 24.8 15.4

BD1: Frequencies for knh20_2$BD1
        2    1   NA
     3641  657  190
%    81.1  14.6 4.2
%!NA 84.7  15.3

HE_HCHOL: Frequencies for knh20_2$HE_HCHOL
        0    1   NA
     2747  1368  373
%    61.2  30.5   8.3
%!NA 66.8  33.2
```

2. 1번의 40세 이상 데이터셋(n=4,488 확인)에 대하여 연령그룹에 따른 총콜레스테롤(HE_chol)의 평균차이를 분산분석으로 검정하고자 한다.

2-1. 연령그룹변수(AGEGP: 1. 40대, 2. 50대, 3. 60대, 4. 70대 이상)을 생성하고 각 범주의 빈도와 %를 제시하시오.

- **Script:**

```
knh20_1$AGEGP=as.factor(ifelse(knh20_1$age<50,'1',ifelse(knh20_1$age<60,'2',ifelse(knh20_1$age<70,'3',ifelse(knh20_1$age,'4')))))
freq(knh20_1$AGEGP)
```

- **설명:** 연령그룹변수 AGEGP를 knh20_1 데이터셋에 추가하였다. 이후 해당 변수의 빈도를 확인하였다.

- **AGEGP 변수의 빈도**: Frequencies for knh20_1$AGEGP

	4	3	2	1	NA
	1212	1148	1106	1022	0
%	27	25.6	24.6	22.8	0
%!NA	27	25.6	24.6	22.8	

2-2. 연령그룹별로 총콜레스테롤(HE_chol) 변수가 정규분포한다고 할 수 있는지 검정하고 등분산 가정이 성립하는지 알아보시오 (귀무가설 제시).

- **Script:**

```
shapiro.test(knh20_1$HE_chol[knh20_1$AGEGP==1])
shapiro.test(knh20_1$HE_chol[knh20_1$AGEGP==2])
shapiro.test(knh20_1$HE_chol[knh20_1$AGEGP==3])
shapiro.test(knh20_1$HE_chol[knh20_1$AGEGP==4])
boxplot(HE_chol~AGEGP,data=knh20_1)
bartlett.test(HE_chol~AGEGP,data=knh20_1)
```

- **설명:**

정규성 검정과 3군 이상 분산 비교를 실행하였다. 정규성 검정의 귀무가설은 "연령 그룹 별 총콜레스테롤 변수는 정규분포를 따른다" 이었고, 모든 연령 그룹에서 귀무가설은

기각되었다. 이때 p-value는 40대 9.939e-06, 50대 2.841e-07, 60대 2.176e-05, 70대 이상 4.001e-06이 도출되었다. 이를 통해 모든 연령대의 총콜레스테롤 변수가 정규성을 띄지 않음을 알 수 있다.

다음으로 분산 비교를 위해 박스 그림을 그리고 3군 이상 분산 비교 분석을 진행했다. 이때 귀무가설은 "모든 분산이 동일하다"였고, 검정 결과 p-value 0.0003252로 귀무가설이 기각되었다. 이를 통해 모든 연령대에서 총콜레스테롤 변수의 모분산이 전부 같지는 않다는 결론을 내릴 수 있다. 또한, 박스그림을 통해서도 네가지 연령군에서의 분산에 차이가 있음을 유추할 수 있다.

- **40대 총콜레스테롤 정규성 검정:**

```
Shapiro-Wilk normality test
data:  knh20_1$HE_chol[knh20_1$AGEGP == 1]
W = 0.99074, p-value = 9.939e-06
```

- **50대 총콜레스테롤 정규성 검정:**

```
      Shapiro-Wilk normality test
data:  knh20_1$HE_chol[knh20_1$AGEGP == 2]
W = 0.98857, p-value = 2.841e-07
```

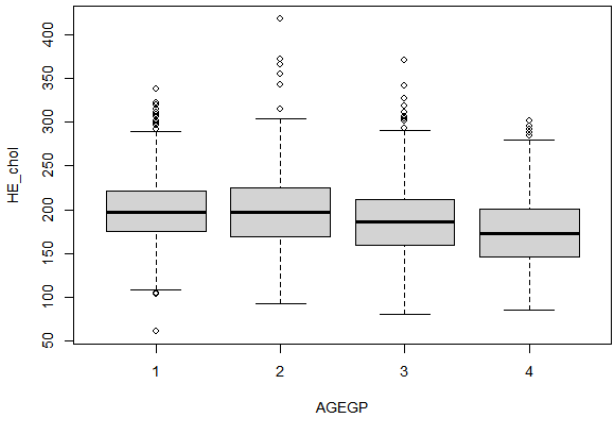
- **60대 총콜레스테롤 정규성 검정:**

```
      Shapiro-Wilk normality test
data:  knh20_1$HE_chol[knh20_1$AGEGP == 3]
W = 0.99247, p-value = 2.176e-05
```

- **70대 총콜레스테롤 정규성 검정:**

```
      Shapiro-Wilk normality test
data:  knh20_1$HE_chol[knh20_1$AGEGP == 4]
W = 0.99152, p-value = 4.001e-06
```

- **박스그림:**



- **분산비교:**

```
Bartlett test of homogeneity of variances
data:  HE_chol by AGEGP
Bartlett's K-squared = 18.635, df = 3, p-value = 0.0003252
```

2-3. 연령그룹에 따른 총콜레스테롤(HE_chol)의 평균 차이를 분산분석으로 검정하고 해석하시오. (2-2의 결과에 무관하게 ANOVA실시, 귀무가설 제시)

- **Script:**

knh20_HE_chol_anova=aov(HE_chol~as.factor(AGEGP),data=knh20_1)

summary(knh20_HE_chol_anova)

- **설명**: 연령 그룹에 따른 총콜레스테롤 평균의 차이를 분산분석으로 검정하였다. 이때 귀무가설은 "모든 연령 그룹의 총콜레스테롤 평균이 같다"이었으며, 검정 결과 p-value는 2e-16 미만으로 귀무가설이 기각되었다. 즉, 연령 그룹에 따른 총콜레스테롤 중 적어도 한 쌍의 평균이 서로 다르다는 것을 알 수 있다.

- **분산분석 결과:**

```
                  Df  Sum Sq  Mean Sq  F value  Pr(>F)
as.factor(AGEGP)  3   398174  132725    85.35   <2e-16 ***
Residuals        4236 6587542  1555
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
결측으로 인하여 248개의 관측치가 삭제되었습니다.
```

2-4. TukeyHSD 법으로 사후분석을 실시하고 결과를 해석하시오.

- **Script:**

TukeyHSD(knh20_HE_chol_anova,'as.factor(AGEGP)',conf.level=0.95)

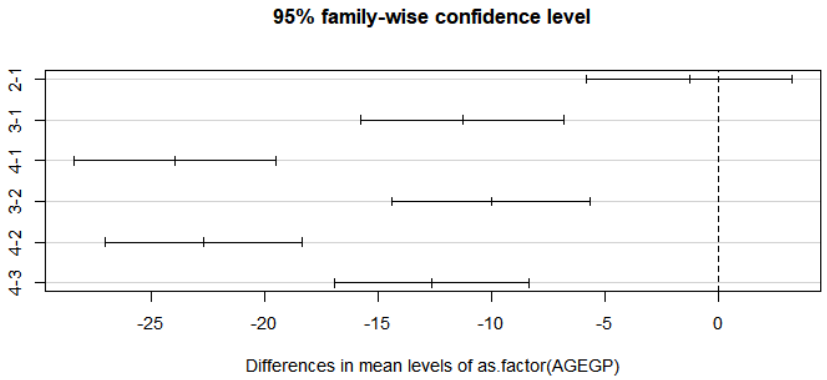
plot(TukeyHSD(knh20_HE_chol_anova,'as.factor(AGEGP)',conf.level=0.95))

- **설명:** 사후 검정을 실행하였고, 신뢰구간의 그래프 또한 그려 보았다. 그 결과 연령 그룹 1과 2 평균 차이의 신뢰구간에 0이 포함된다는 사실을 발견하였다. 1과 2 사이를 제외한 다른 신뢰구간은 모두 0을 포함하지 않았다. 이를 통해, 40대와 50대 사이의 총콜레스테롤 수치를 제외한 다른 연령 그룹 별 평균이 모두 다르다는 해석을 내릴 수 있다.

- **TukeyHSD 사후분석 결과**

```
Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = HE_chol ~ as.factor(AGEGP), data = knh20_1)
$`as.factor(AGEGP)`
          diff        lwr        upr     p adj
2-1  -1.276684  -5.809059   3.255692 0.8875715
3-1 -11.291163 -15.770990  -6.811337 0.0000000
4-1 -23.954951 -28.399776 -19.510127 0.0000000
3-2 -10.014480 -14.392468  -5.636491 0.0000001
4-2 -22.678268 -27.020434 -18.336102 0.0000000
4-3 -12.663788 -16.951074  -8.376502 0.0000000
```

- **사후분석 신뢰구간 그래프:**



3. 연령이 40대인 대상자에 한하여(n=1,022 확인) 총콜레스테롤(HE_chol)과 허리둘레(HE_wc)의 상관관계를 평가하고자 한다.

3-1. 두 변수가 정규분포한다고 할 수 있는지 검정하시오 (귀무가설 제시).

- **Script:**

  knh20_40=subset(knh20_1,knh20_1$AGEGP==1)

  shapiro.test(knh20_40$HE_chol)

  shapiro.test(knh20_40$HE_wc)

- **설명**: 40대 대상자만을 포함하는 새로운 데이터셋 "knh20_40"을 생성하였다. 이후 총콜레스테롤과 허리둘레 변수의 정규성을 각각 검정하였다. 검정 시 귀무가설은 "해당 변수가 정규 분포를 따른다" 이었으며, 검정 결과 두 변수 모두 귀무가설이 기각되었고, 모두 정규성을 띄지 않는다는 점을 알 수 있다. 이때 p-value는 총콜레스테롤 9.939e-06, 허리둘레 3.466e-07이 도출되었다.

- **N=1022 확인**

- **HE_chol 정규성 검정**:

```
Shapiro-Wilk normality test
data:  knh20_40$HE_chol
W = 0.99074, p-value = 9.939e-06
```

- **HE_wc 정규성 검정:**

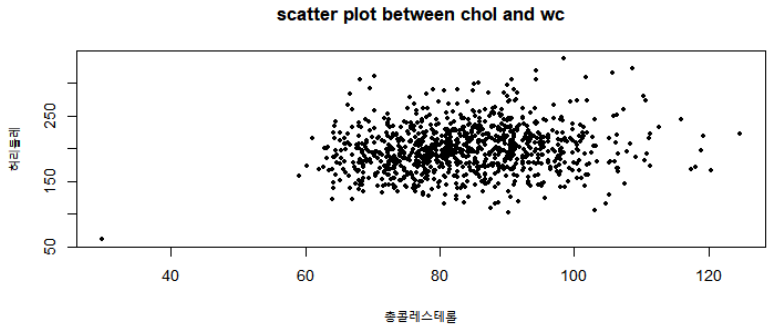
```
Shapiro-Wilk normality test
data:  knh20_40$HE_wc
W = 0.98785, p-value = 3.466e-07
```

3-2. 산점도를 그리시오.

- **Script:**

plot(HE_chol~HE_wc,data=knh20_40,xlab='총콜레스테롤',ylab="허리둘레",main='scatter plot between chol and wc',type='p',pch=20,cex=1)

- **산점도 그래프:**



3-3. 피어슨 및 스피어만 상관계수를 구하고 유의성을 검정하시오 (귀무가설 제시)

- **Script:**

cor(knh20_40[c('HE_chol','HE_wc')],use='complete.obs',method='pearson')

cor(knh20_40[c('HE_chol','HE_wc')],use='complete.obs',method='spearman')

cor.test(~HE_chol+HE_wc,data=knh20_40,method=c('pearson'))

cor.test(~HE_chol+HE_wc,data=knh20_40,method=c('spearman'))

- **설명:** 총콜레스테롤 변수와 허리둘레 변수 사이의 피어슨과 스페어만 상관계수를 각각 구하였고 Pearson 상관 계수는 0.1514348, spearman 상관계수는 0.1451427이 도출되었다. 이후 유의성을 검정하였다. 유의성 검정 시 귀무가설은 "상관계수가 0이다"이었고, 검정 결과 p-value 2.623e-06로 귀무가설이 기각되었다. 이는 총콜레스테롤과 허리둘레 변수의 상관계수가 0이 아니라는 것을 의미하며, 둘 사이의 상관성이 존재함을 보여준다.

- **Pearson 상관계수:**

```
HE_chol      HE_wc
HE_chol 1.0000000 0.1514348
HE_wc   0.1514348 1.0000000
```

- **Spearman 상관계수:**

```
HE_chol      HE_wc
```

```
HE_chol 1.0000000 0.1451427
HE_wc   0.1451427 1.0000000
```

- **Pearson 유의성 검정:**

```
Pearson's product-moment correlation
data:  HE_chol and HE_wc
t = 4.727, df = 952, p-value = 2.623e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.08881776 0.21285963
sample estimates:
      cor
0.1514348
```

- **Spearman 유의성 검정:**

```
Spearman's rank correlation rho
data:  HE_chol and HE_wc
S = 123704935, p-value = 6.764e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1451427
```

4. 총콜레스테롤(HE_chol)을 설명하는 회귀분석을 실시하시오
(40세 이상인 대상자로 AGEGP, age, sex, HE_wt, HE_ht, HE_wc, smoke, BD1, HE_chol를 포함하는 자료로 결측치가 모두 제거된(na.omit 이용) 데이터셋을 추출하여 4번 문제에 사용하시오. n=4,118 확인)

4-1. BMI 생성: 체중(HE_wt)과 키(HE_ht/100)를 이용하여 체질량지수(kg/m^2)로 BMI라는 변수를 생성하고 기술통계를 n, 평균, SD, 최소값, 최대값으로 제시하시오.

- **Script:**

```
knh20_chol=na.omit(subset(knh20_1,select=c('AGEGP','age','sex','HE_wt','HE_ht','HE_wc','smoke','BD1','HE_chol')))
knh20_chol$BMI=(knh20_chol$HE_wt)/((knh20_chol$HE_ht/100)^2)
describe(knh20_chol$BMI)
```

- **설명:** 40세 이상 대상자에서 AGEGP, age, sex, HE_wt, HE_ht, HE_wc. Smoke, BD1, HE_chol 변수를 포함하는 "knh20_chol" 데이터셋을 생성하였다. 이후 체질량지수로 BMI라는 변수를 추가하였고, 기술통계량을 구하였다.

- **N=4118 확인**

- **BMI 변수의 기술통계량:**

	N	평균	SD	최솟값	최댓값
BMI	4118	24.33	5.02	14.16	264.05

4-2. 총콜레스테롤(HE_chol), 연령(age), 성별(sex), BMI, 허리둘레(HE_wc), 흡연상태(smoke), 음주(BD1) 에 대한 이변량 피어슨 및 스피어만 상관계수를 구하고 산점도를 그리시오.

- **Script:**

```
library(Hmisc)
```

```
knh20_matrix=as.matrix(knh20_chol[c('HE_chol','age','sex','BMI','HE_wc','smoke','BD1')])
rcorr(knh20_matrix,type='pearson')
rcorr(knh20_matrix,type='spearman')

library(corrplot)
knh20_chol$smoke=as.numeric(knh20_chol$smoke)
knh20_chol$BD1=as.numeric(knh20_chol$BD1)

knh20_cor_p=cor(knh20_chol[c('HE_chol','age','sex','BMI','HE_wc','smoke','BD1')],use='complete.obs',method=c('pearson'))
knh20_cor_p
knh20_cor_s=cor(knh20_chol[c('HE_chol','age','sex','BMI','HE_wc','smoke','BD1')],use='complete.obs',method=c('spearman'))
knh20_cor_s

plot(knh20_chol[c('HE_chol','age','sex','BMI','HE_wc','smoke','BD1')],pch=8,cex=0.5,ellipse=T)
```

- **설명**: 변수들을 matrix 형태로 변환하고 이에 대한 이변량 피어슨 및 스피어만 상관계수를 구하였다. 그 결과 아래와 같은 결과가 도출되었다. 또한, 더욱 구체적인 상관계수 수치를 구하기 위해 smoke 변수와 BD1 변수를 수치형으로 변환하고 상관계수를 구하였다. 이후 각 변수에 대한 산점도를 그렸다.

- **Pearson 상관계수:**

```
        HE_chol   age    sex    BMI HE_wc smoke   BD1
HE_chol    1.00 -0.22  0.10 -0.04 -0.08 -0.04  0.05
age       -0.22  1.00  0.00 -0.05  0.14 -0.11 -0.28
sex        0.10  0.00  1.00 -0.06 -0.34 -0.66 -0.25
BMI       -0.04 -0.05 -0.06  1.00  0.51  0.05  0.02
HE_wc     -0.08  0.14 -0.34  0.51  1.00  0.19  0.03
smoke     -0.04 -0.11 -0.66  0.05  0.19  1.00  0.24
BD1        0.05 -0.28 -0.25  0.02  0.03  0.24  1.00

n= 4118

P
        HE_chol age    sex    BMI    HE_wc  smoke  BD1
HE_chol         0.0000 0.0000 0.0077 0.0000 0.0131 0.0014
age     0.0000         0.8537 0.0038 0.0000 0.0000 0.0000
```

sex	0.0000	0.8537		0.0000	0.0000	0.0000
BMI	0.0077	0.0038	0.0000		0.0000	0.0017	0.1652
HE_wc	0.0000	0.0000	0.0000	0.0000		0.0000	0.0340
smoke	0.0131	0.0000	0.0000	0.0017	0.0000		0.0000
BD1	0.0014	0.0000	0.0000	0.1652	0.0340	0.0000	

- **Spearman 상관계수:**

	HE_chol	age	sex	BMI	HE_wc	smoke	BD1
HE_chol	1.00	-0.23	0.10	-0.01	-0.09	-0.05	0.05
age	-0.23	1.00	0.00	-0.01	0.15	-0.09	-0.28
sex	0.10	0.00	1.00	-0.13	-0.35	-0.70	-0.25
BMI	-0.01	-0.01	-0.13	1.00	0.85	0.07	0.03
HE_wc	-0.09	0.15	-0.35	0.85	1.00	0.22	0.03
smoke	-0.05	-0.09	-0.70	0.07	0.22	1.00	0.26
BD1	0.05	-0.28	-0.25	0.03	0.03	0.26	1.00

n= 4118

P

	HE_chol	age	sex	BMI	HE_wc	smoke	BD1
HE_chol		0.0000	0.0000	0.3763	0.0000	0.0018	0.0017
age	0.0000		0.8506	0.3696	0.0000	0.0000	0.0000
sex	0.0000	0.8506		0.0000	0.0000	0.0000	0.0000
BMI	0.3763	0.3696	0.0000		0.0000	0.0000	0.0804
HE_wc	0.0000	0.0000	0.0000	0.0000		0.0000	0.0354
smoke	0.0018	0.0000	0.0000	0.0000	0.0000		0.0000
BD1	0.0017	0.0000	0.0000	0.0804	0.0354	0.0000	

- **구체적 수치 (pearson 상관계수)**

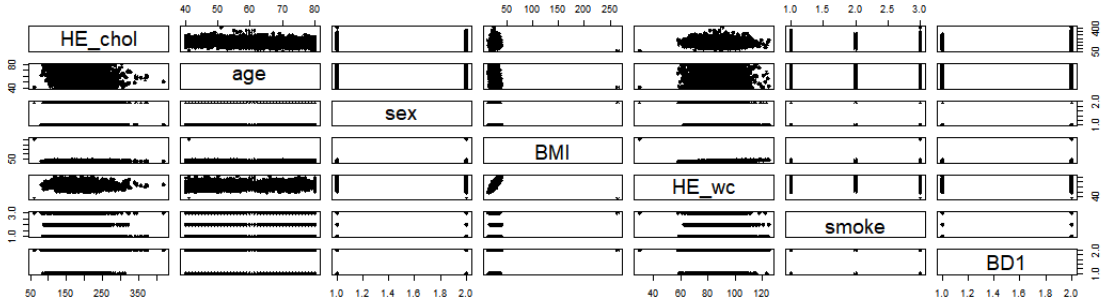
HE_chol	age	sex	BMI	HE_wc	smoke	BD1
HE_chol 1.00000000	-0.223249671	0.103838876	-0.04152833	-0.07901475	-0.03868025	0.04972501
age -0.22324967	1.000000000	0.002874105	-0.04515150	0.13851742	-0.11116885	-0.27790412
sex 0.10383888	0.002874105	1.000000000	-0.06142689	-0.33506782	-0.65757516	-0.24908656
BMI -0.04152833	-0.045151505	-0.061426889	1.00000000	0.50711116	0.04892195	0.02162938
HE_wc -0.07901475	0.138517421	-0.335067821	0.50711116	1.00000000	0.18634647	0.03303415
smoke -0.03868025	-0.111168852	-0.657575157	0.04892195	0.18634647	1.00000000	0.24089447
BD1 0.04972501	-0.277904118	-0.249086561	0.02162938	0.03303415	0.24089447	1.00000000

- **구체적 수치 (spearman 상관계수)**

HE_chol	age	sex	BMI	HE_wc	smoke	BD1
HE_chol 1.00000000	-0.227818459	0.101637898	-0.01379053	-0.09174979	-0.04850344	0.04897082
age -0.22781846	1.000000000	0.002936742	-0.01398418	0.14729888	-0.09325476	-0.27747615
sex 0.10163790	0.002936742	1.000000000	-0.12917078	-0.34676784	-0.70381729	-0.24908656
BMI -0.01379053	-0.013984179	-0.129170780	1.00000000	0.85058375	0.06787002	0.02724407

HE_wc	-0.09174979	0.147298880	-0.346767844	0.85058375	1.00000000	0.22143606	0.03278589
smoke	-0.04850344	-0.093254757	-0.703817292	0.06787002	0.22143606	1.00000000	0.25666546
BD1	0.04897082	-0.277476150	-0.249086561	0.02724407	0.03278589	0.25666546	1.00000000

- **산점도:**



4-3. 연령군 변수를 이용하여 총콜레스테롤을 설명하는 단순회귀분석을 실시하시오.

4-3-1. 모형의 적합도를 F통계량을 이용하여 평가하시오 (귀무가설 제시). 또한 R-square를 구하고 의미를 설명하시오.

- **script:**

```
model_chol=lm(HE_chol~AGEGP,data=knh20_chol)
print(model_chol)
summary(model_chol)
anova(model_chol)
```

- **설명:**

총콜레스테롤 변수가 종속변수, 연령군 변수가 독립변수인 회귀모형 'model_chol'을 생성하였다. 이때 독립 변수인 연령군이 범주형 변수임에 따라 가변수인 AGEGP2, AGEGP3, AGEGP4가 생성되었다. F통계량을 통해 모형의 적합도를 판단하였으며 이때 귀무가설은 "회귀계수 값이 0이다"이다. 평가 결과 귀무가설은 기각되었으며, 이는 해당 회귀계수가 독립변수와 종속변수의 관계를 설명하기에 적합하다는 것을 의미한다.

R-square는 결정계수로, 종속변수의 전체 변동 중 회귀모형에 의해 설명되는 변동의 크기를 나타낸다. 이는 회귀식에 의해 예측된 종속변수의 값과 실제로 관찰된 종속변수 사이의 상관계수의 제곱이며, 결론적으로 회귀모형의 설명력을 의미한다. model_chol의 multiple R-square 값은 0.05545, adjusted R-squared 값은 0.05476이 도출되었다. 즉,

약 5%의 총콜레스테롤 수치의 변동이 해당 회귀모형으로 설명된다는 것이다.

- **회귀모형:**

```
Call:
lm(formula = HE_chol ~ AGEGP, data = knh20_chol)
Coefficients:
(Intercept)       AGEGP2       AGEGP3       AGEGP4
    198.781       -1.453      -11.569      -23.890
```

- **회귀모형 요약 (r-square):**

```
Call:
lm(formula = HE_chol ~ AGEGP, data = knh20_chol)
Residuals:
     Min       1Q   Median       3Q      Max
-137.781  -27.329   -1.212   25.108  220.671

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  198.781      1.284 154.776  < 2e-16 ***
AGEGP2        -1.453      1.776  -0.818    0.413
AGEGP3       -11.569      1.762  -6.567 5.79e-11 ***
AGEGP4       -23.890      1.766 -13.525  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.5 on 4114 degrees of freedom
Multiple R-squared:  0.05545,    Adjusted R-squared:  0.05476
F-statistic: 80.51 on 3 and 4114 DF,  p-value: < 2.2e-16
```

- **모형 적합도 검정:**

```
Analysis of Variance Table
Response: HE_chol
            Df  Sum Sq Mean Sq F value    Pr(>F)
AGEGP        3  376874  125625  80.509 < 2.2e-16 ***
Residuals 4114 6419439    1560
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4-3-2. 회귀계수의 유의성 검정을 실시하고(귀무가설 제시), 분석 결과를 해석하시오.

- **Script:**

  summary(model_chol)

  설명: 회귀계수의 유의성 검정은 t-test를 통해 진행된다. 이때 귀무가설은 "각 연령그룹 별 회귀계수가 0이다"이며, 검정 결과 AGEGP2를 제외한 가변수에서 귀무가설은 기각되었다. 이를 통해 AGEGP2의 회귀계수가 0이 된다는 것을 알 수 있다. 즉, 40대와 50대 총콜레스테롤의 기댓값에 차이가 존재하고, 나머지 연령대 사이에는 유의미한 차이가 없다는 것이다.

- **회귀모형 요약 (유의성 t-test):**

```
Call:
lm(formula = HE_chol ~ AGEGP, data = knh20_chol)
Residuals:
     Min       1Q   Median       3Q      Max
-137.781  -27.329   -1.212   25.108  220.671

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  198.781      1.284 154.776  < 2e-16 ***
AGEGP2        -1.453      1.776  -0.818    0.413
AGEGP3       -11.569      1.762  -6.567 5.79e-11 ***
AGEGP4       -23.890      1.766 -13.525  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.5 on 4114 degrees of freedom
Multiple R-squared:  0.05545,    Adjusted R-squared:  0.05476
F-statistic: 80.51 on 3 and 4114 DF,  p-value: < 2.2e-16
```

4-3-3. 잔차그림(Residual Plot)을 그리고 회귀모형의 선형성과 등분산성 및 잔차의 정규성 가정에 적합한 지에 대하여 논하시오.

- **Script:**

par(mfrow=c(2,2))

plot(model_chol)

resid_chol=resid(model_chol)

stdres_chol=rstandard(model_chol)

pred_chol=predict(model_chol)

par(mfrow=c(1,1))

hist(stdres_chol)

library(ggplot2)

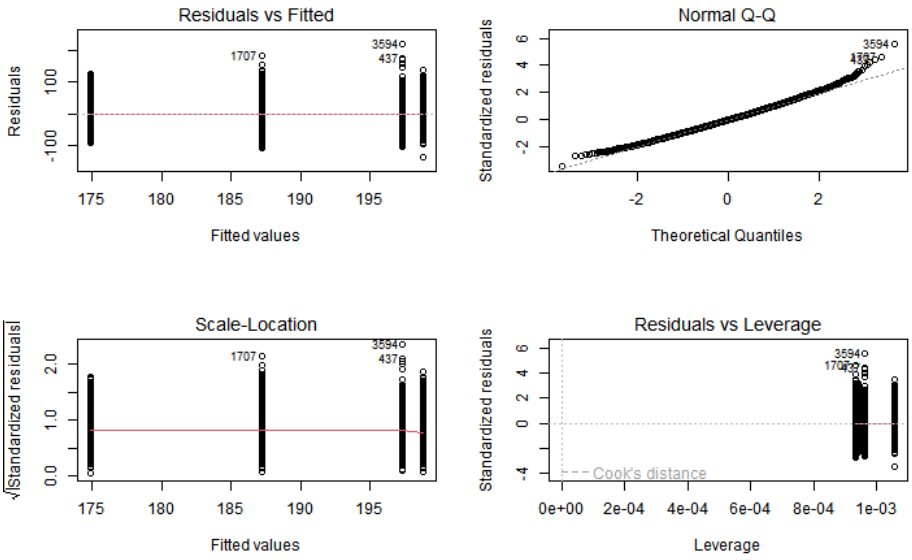
ggplot(knh20_chol)+geom_point(mapping=aes(x=AGEGP,y=resid_chol))+geom_hline(yintercept=0, linetype='dashed', color='red', size=0.5)

- **설명:**

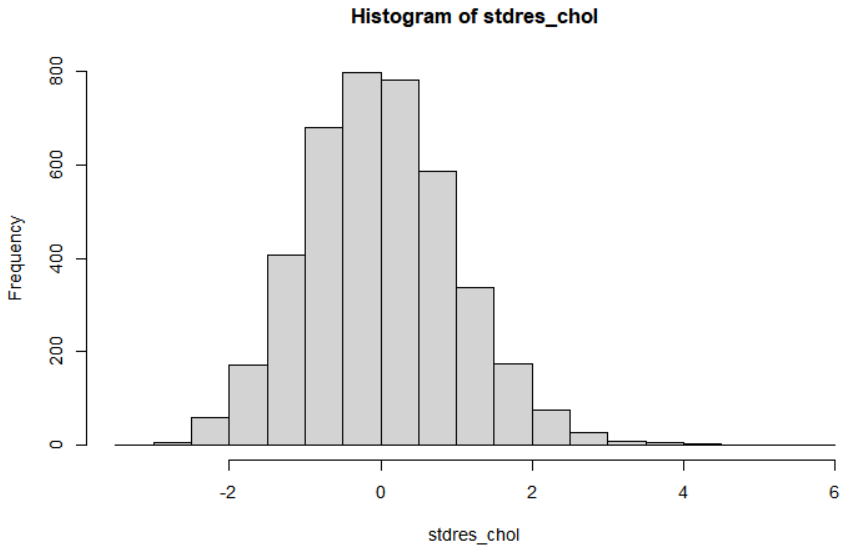
먼저 plot 창을 2:2로 분할 후 회귀모형 model_chol에 대한 잔차 plot을 그렸다. 이때 왼쪽 위의 그래프를 통해 등분산성을 확인할 수 있었고, 오른쪽 위 그래프를 통해 잔차의 정규성을 확인할 수 있었다. 확인 결과 해당 모형이 등분산성을 만족하고, 정규 분포를 따른다는 사실을 알 수 있었다.

이후 분할된 창을 원래대로 복구하고 histogram of stadardized residual 그래프를 그렸고, residual*X 그래프를 통해 선형성을 알아볼 수 있었다. 분석 결과 해당 회귀모형은 앞선 결과처럼 정규성을 만족하는 히스토그램과 유사한 형태로 그려졌고, 선형성도 만족된다는 것을 알 수 있었다.

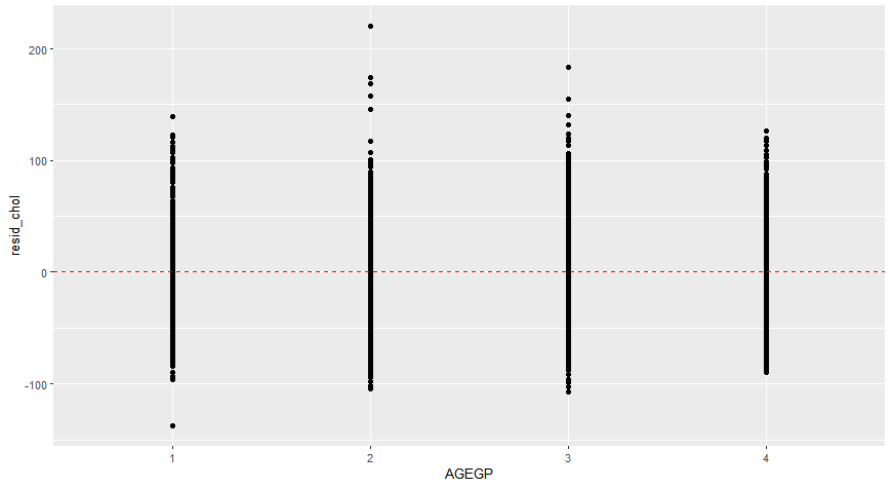
- **잔차 그림**



- **잔차의 histogram**



- **Linearity assumption**



4-4. 연령군(AGEGP), 성별(sex), BMI, 허리둘레(HE_wc), 흡연상태(smoke), 음주(BD1)를 독립변수로 포함하는 다중회귀모형을 분석하시오.

- **Script:**

model_chol_1=lm(HE_chol~AGEGP+factor(sex)+BMI+HE_wc+factor(smoke)+factor(BD1),dat

a=knh20_chol)

- **설명:** 연령군, 성별, BMI, 허리둘레, 흡연상태, 음주를 독립변수로 하고 총콜레스테롤 수치를 종속변수로 하는 새로운 회귀모형 'model_chol_1'을 생성하였다. 이때 독립변수 중 범주형변수인 AGEGP, sex, smoke, BD1에서 각각 가변수들이 형성되었다.

4-4-1. 다중공선성이란 무엇인지 설명하고 현재 모형에서의 VIF에 대한 판단은 어떠한가?

- **Script:**

install.packages('olsrr')

library(olsrr)

ols_vif_tol(model_chol_1)

- **설명:**

다중공선성이란, 독립변수가 여러 개인 다중 회귀 분석에서 독립변수들 사이에 강한 상관관계가 있는 경우를 의미한다. 이 경우 추정값의 표준오차가 커져 추정된 회귀계수의 신뢰도가 떨어진다는 한계가 있다. VIF (Variance Inflation Factor)인 분산팽창요인이 10 이상인 경우 이러한 다중공선성이 있다고 판단된다.

R에서 olsrr 패키지를 설치 및 실행 후 회귀모형의 분산팽창요인을 알아보았고, 분석결과 VIF 값이 10을 초과하는 변수는 존재하지 않았다. 이를 통해 해당 회귀모형이 다중공선성 문제를 지니지 않음을 알 수 있다.

- **VIF:**

	Variables	Tolerance	VIF
1	AGEGP2	0.6329301	1.579953
2	AGEGP3	0.6112527	1.635985
3	AGEGP4	0.5635698	1.774403
4	factor(sex)2	0.4267995	2.343021
5	BMI	0.7162119	1.396235
6	HE_wc	0.6232353	1.604530
7	factor(smoke)2	0.4985302	2.005897
8	factor(smoke)3	0.5887465	1.698524
9	factor(BD1)2	0.8482731	1.178866

4-4-2. 모형의 적합도를 F통계량을 이용하여 평가하시오 (귀무가설 제시). 또한 R-square 결과의 의미를 설명하시오.

- **Script:**

  anova(model_chol_1)

  summary(model_chol_1)

- **설명**:

  모형의 적합도를 F-통계량을 통해 확인하였다. 이때 귀무가설은 "모든 회귀계수 값이 0이다"이다. 평가 결과 귀무가설이 기각되었다. 이는 모든 회귀계수의 값이 0이 맞으며, 해당 회귀모형이 적합함을 의미한다.

  다음으로 model_chol_1의 Multiple R-squared 값은 0.06907, Adjusted R-squared 값은 0.06703 이 도출되었다. 이를 통해 약 7%의 총콜레스테롤 수치의 변동이 해당 회귀모형으로 설명된다는 것으로 해석할 수 있다.

- **회귀모형 요약 (r-square, 적합도 검정):**

```
Call:
lm(formula = HE_chol ~ AGEGP + factor(sex) + BMI + HE_wc + factor(smoke) + factor(BD1),
data = knh20_chol)
Residuals:
    Min      1Q  Median      3Q     Max
-107.14  -27.11   -1.37   24.38  223.51
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    197.72102    6.53625  30.250  < 2e-16 ***
AGEGP2          -1.44642    1.77040  -0.817  0.41398
AGEGP3         -11.57336    1.78205  -6.494 9.33e-11 ***
AGEGP4         -23.64769    1.86271 -12.695  < 2e-16 ***
factor(sex)2     8.24303    1.88550   4.372 1.26e-05 ***
BMI             -0.41031    0.14394  -2.851  0.00439 **
HE_wc            0.06379    0.07910   0.806  0.42003
factor(smoke)2  -2.18877    2.00781  -1.090  0.27572
factor(smoke)3   1.08085    2.20839   0.489  0.62456
factor(BD1)2     1.49711    1.87304   0.799  0.42417
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 39.24 on 4108 degrees of freedom
Multiple R-squared:  0.06907,    Adjusted R-squared:  0.06703
F-statistic: 33.86 on 9 and 4108 DF,  p-value: < 2.2e-16
```

4-4-3. 여러 회귀계수의 유의성 검정을 실시하고 분석결과를 해석하시오.

- **Script:**

  summary(model_chol_1)

- **설명:** 회귀계수의 유의성을 t-test를 통해 분석하였다. 이 때 귀무가설은 '각 회귀계수는 0이다' 이었으며, 분석 결과 AGEGP2, HE_wc, smoke2, smoke3, BD1 2 변수에서 귀무가설이 수용되었음을 확인하였다. 즉, 해당 변수에서의 회귀계수가 0이며, 그에 따라 회귀계수가 유의하지 않을 수 있다는 것이다. 또한 이중 허리둘레를 제외한 독립변수들은 범주형 변수에서 기인한 가변수들이다. 따라서 AGEGP2, smoke2, smoke3, BD1 2의 회귀계수가 0이라는 것은 해당 변수들과 기준 변수의 기댓값에 유의미한 차이가 존재한다는 것을 의미한다.

- **회귀모형 요약 (유의성 t-test):**

```
Call:
lm(formula = HE_chol ~ AGEGP + factor(sex) + BMI + HE_wc + factor(smoke) + factor(BD1),
data = knh20_chol)
Residuals:
    Min      1Q  Median      3Q     Max
-107.14  -27.11   -1.37   24.38  223.51
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    197.72102    6.53625  30.250  < 2e-16 ***
AGEGP2          -1.44642    1.77040  -0.817  0.41398
AGEGP3         -11.57336    1.78205  -6.494 9.33e-11 ***
AGEGP4         -23.64769    1.86271 -12.695  < 2e-16 ***
factor(sex)2     8.24303    1.88550   4.372 1.26e-05 ***
BMI             -0.41031    0.14394  -2.851  0.00439 **
HE_wc            0.06379    0.07910   0.806  0.42003
factor(smoke)2  -2.18877    2.00781  -1.090  0.27572
factor(smoke)3   1.08085    2.20839   0.489  0.62456
factor(BD1)2     1.49711    1.87304   0.799  0.42417
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 39.24 on 4108 degrees of freedom
Multiple R-squared:  0.06907,    Adjusted R-squared:  0.06703
F-statistic: 33.86 on 9 and 4108 DF,  p-value: < 2.2e-16
```

4-4-4. 표준화회귀계수를 구하고 독립변수인 연령군(AGEGP), 성별(sex), BMI, 허리둘레(HE_wc), 흡연상태(smoke), 음주(BD1)의 중요도를 비교하시오.

- **Script:**

  library(lm.beta)

  model_chol.beta=lm.beta(model_chol_1)

  print(model_chol.beta)

- **설명**: lm.beta를 사용하여 각 독립변수 간의 표준화회귀계수를 구하였다. 분석 결과 아래와 같은 수치들이 도출되었다. 표준화회귀계수의 절댓값이 클수록 독립변수가 종속변수에 미치는 영향의 크기가 크기 때문에 이를 통해 독립변수 사이의 중요도를 비교할 수 있다. 표준화 회귀계수의 절댓값은 AGEGP4, AGEGP3, sex2, BMI, smoke2, AGEGP2, HE_wc, BD1 2, smoke3 순서대로 작아진다. 즉, 70세 이상 연령군의 가변수의 중요도가 가장 크고, 현재 흡연 가변수의 중요도가 가장 작음을 알 수 있다.

- **표준화 회귀계수:**

```
Call:
lm(formula = HE_chol ~ AGEGP + factor(sex) + BMI + HE_wc + factor(smoke) + factor(BD1),
data = knh20_chol)

Standardized Coefficients::
  (Intercept)         AGEGP2         AGEGP3         AGEGP4   factor(sex)2
           NA   -0.015459238   -0.125046952   -0.254574150    0.100737727

          BMI         HE_wc factor(smoke)2 factor(smoke)3   factor(BD1)2
-0.050705810    0.015377887   -0.023242046    0.009602166    0.013064181
```

4-5. 4-3) 단순회귀모형과 4-4) 다중회귀모형을 비교하고 더 좋은 설명력을 보이는 모형을 선택하시오.

4-5-1 수정결정계수를 이용하여 더 좋은 설명력을 보이는 모형을 선택하시오.

- **Script:**

  summary(model_chol)

  summary(model_chol_1)

- **설명:** model_chol과 model_chol_1의 adjusted r-square 값을 비교하였다. 비교 결과 model_chol의 수정결정계수는 0.05476, model_chol_1의 수정결정계수는 0.06703이 도출되었다. 즉, model_chol은 총콜레스테롤 변화의 5.476%를, model_chol_1은 6.703%를 설명해주기 때문에 후자의 설명력이 더 좋다고 결론지을 수 있다.

- **model_chol 요약 (adj r-square)**

```
Call:
lm(formula = HE_chol ~ AGEGP, data = knh20_chol)

Residuals:
     Min       1Q   Median
-137.781  -27.329   -1.212
      3Q      Max
  25.108  220.671

Coefficients:
            Estimate Std. Error
(Intercept)  198.781      1.284
AGEGP2        -1.453      1.776
AGEGP3       -11.569      1.762
AGEGP4       -23.890      1.766
            t value Pr(>|t|)
(Intercept) 154.776  < 2e-16 ***
AGEGP2       -0.818    0.413
AGEGP3       -6.567 5.79e-11 ***
AGEGP4      -13.525  < 2e-16 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01
   ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 39.5 on 4114 degrees of freedom
Multiple R-squared:  0.05545,    Adjusted R-squared:  0.05476
F-statistic: 80.51 on 3 and 4114 DF,  p-value: < 2.2e-16
```

- **model_chol_1 요약 (adj r-square)**

```
Call:
lm(formula = HE_chol ~ AGEGP + factor(sex) + BMI + HE_wc + factor(smoke) +
    factor(BD1), data = knh20_chol)

Residuals:
    Min      1Q  Median      3Q
-107.14  -27.11   -1.37   24.38
    Max
 223.51

Coefficients:
               Estimate
(Intercept)    197.72102
AGEGP2          -1.44642
AGEGP3         -11.57336
AGEGP4         -23.64769
factor(sex)2     8.24303
BMI             -0.41031
HE_wc            0.06379
factor(smoke)2  -2.18877
factor(smoke)3   1.08085
factor(BD1)2     1.49711
               Std. Error t value
(Intercept)       6.53625  30.250
AGEGP2            1.77040  -0.817
AGEGP3            1.78205  -6.494
AGEGP4            1.86271 -12.695
factor(sex)2      1.88550   4.372
BMI               0.14394  -2.851
HE_wc             0.07910   0.806
factor(smoke)2    2.00781  -1.090
factor(smoke)3    2.20839   0.489
factor(BD1)2      1.87304   0.799
               Pr(>|t|)
(Intercept)     < 2e-16 ***
AGEGP2          0.41398
AGEGP3         9.33e-11 ***
```

```
AGEGP4          < 2e-16 ***
factor(sex)2   1.26e-05 ***
BMI             0.00439 **
HE_wc           0.42003
factor(smoke)2  0.27572
factor(smoke)3  0.62456
factor(BD1)2    0.42417
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01
  ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.24 on 4108 degrees of freedom
Multiple R-squared:  0.06907,    Adjusted R-squared:  0.06703
F-statistic: 33.86 on 9 and 4108 DF,  p-value: < 2.2e-16
```

4-5-2 R-square의 증가량을 이용한 F-test로 비교하여 더 좋은 설명력을 보이는 모형을 선택하시오. (귀무가설 제시)

- **script:**

  anova(model_chol,model_chol_1)

- **설명:** model_chol과 model_chol_1의 설명력을 비교하였다. 이때 귀무가설은 '두 모형의 설명력이 같다'이며, 대립가설은 "복잡한 모형의 설명력이 더 낫다"이다. 검정 결과 p-value 5.235e-11로 귀무가설이 기각되었다. 즉, model_chol_1의 설명력이 더 낫다는 결론이 도출된다.

- **모델 설명력 비교**

```
Analysis of Variance Table
Model 1: HE_chol ~ AGEGP
Model 2: HE_chol ~ AGEGP + factor(sex) + BMI + HE_wc + factor(smoke) + factor(BD1)
Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   4114 6419439
2   4108 6326902  6     92537 10.014 5.235e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5. 고콜레스테롤혈증 유병 여부(HE_HCHOL)를 종속변수로 로지스틱회귀분석을 실시하시오.
(40세 이상인 대상자로 HE_HCHOL, age, sex, HE_wt, HE_ht, HE_wc, smoke, BD1, BMI를 포함하는 자료로 결측치가 모두 제거된(na.omit 이용) 데이터셋을 추출하여 5번 문제에 사용하시오. n=4,006 확인)

5-1. 연령(age) 변수를 독립변수로 포함하는 단순 로지스틱회귀모형을 분석하시오. 연령이 1년 증가하면 콜레스테롤혈증(HE_HCHOL) 유병의 변화는 어떠한지에 대하여 분석결과를 해석하시오.

- **Script:**

```
knh20_hchol=na.omit(subset(knh20_1,select=c('HE_HCHOL','age','sex','HE_wt','HE_ht','HE_wc','smoke','BD1')))
knh20_hchol$BMI=(knh20_hchol$HE_wt)/((knh20_hchol$HE_ht/100)^2)

model_hchol=glm(knh20_hchol$HE_HCHOL~knh20_hchol$age,family=binomial)
summary(model_hchol)
exp(coef(model_hchol))
exp(confint(model_hchol))
```

- **설명:**

우선 관심 변수들을 포함하고 결측치를 제거한 데이터셋 'knh20_hchol'을 생성하고 BMI 변수를 해당 데이터셋에 추가하였다. 이후 고콜레스테롤혈증 유병 여부가 종속변수, 연령이 독립변수인 회귀모형 'model_hchol'을 만들고 단순 로지스틱 회귀분석을 진행하였다. 이후 오즈비인 exp(회귀계수)값을 구하였고, 1.02751146이 도출되었으며, 95% 신뢰구간으로는 [1.02162682, 1.0334676]이 나왔다. 이를 통해 연령 변수가 한 단위 증가할 때 콜레스테롤혈증 유병의 오즈가 1.02751146배 증가한다는 것을 알 수 있다.

- **회귀모형 요약**

```
Call:
glm(formula = knh20_hchol$HE_HCHOL ~ knh20_hchol$age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1055  -0.9230  -0.7871   1.3583   1.7364
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.342797   0.183515 -12.766   <2e-16 ***
knh20_hchol$age  0.027140   0.002939   9.234   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5099.3  on 4005  degrees of freedom
Residual deviance: 5012.1  on 4004  degrees of freedom
AIC: 5016.1

Number of Fisher Scoring iterations: 4
```

- 오즈비:

```
    (Intercept) knh20_hchol$age
     0.09605854      1.02751146
```

- 오즈비 신뢰구간:

```
                       2.5 %    97.5 %
(Intercept)     0.06691232 0.1373978
knh20_hchol$age 1.02162682 1.0334676
```

5-2. 연령(age), 성별(sex), BMI, 허리둘레(HE_wc), 흡연상태(smoke), 음주(BD1)를 독립변수로 포함하는 다중로지스틱회귀모형을 분석하고 그 결과를 연속변수와 범주형 변수로 구분하여 구체적으로 해석하시오.

- **Script:**

```
model_hchol_1=glm(knh20_hchol$HE_HCHOL~knh20_hchol$age+as.factor(knh20_hchol$sex)+knh20_hchol$BMI+knh20_hchol$HE_wc+as.factor(knh20_hchol$smoke)+as.factor(knh20_hchol$BD1),family=binomial)
summary(model_hchol_1)
exp(coef(model_hchol_1))
exp(confint(model_hchol_1))
```

- **설명:**

연령, 성별, BMI, 허리둘레, 흡연상태, 음주 변수를 독립변수로, 고콜레스테롤혈증 유병 여부를 종속변수로 하는 다중 로지스틱 회귀분석 모형을 생성하였다. 분석 결과 연속변수인 연령, BMI, 허리둘레의 오즈비는 각각 1.026072836, 0.997342785, 1.041782556이 도출되었다. 이는 연령, BMI, 허리둘레가 각각 한 단위 증가할 때 고콜레스테롤혈증이 유병될 오즈가 해당 수치만큼 증가함을 의미한다. 또한, 범주형 변수인 sex2, smoke2, smoke3, BD1 2의 오즈비는 각각 2.317667925, 1.298403019, 1.595638129, 1.121126861이 도출되었다. 이는 각각 남성에 비해 여성에게 유병될 오즈, 과거 흡연자에게 유병될 오즈, 현재 흡연자에게 유병될 오즈, 평생 음주 경험이 있는 사람에게 유병될 오즈가 다른 범주에 비해 더 높은 비율을 의미한다.

- **회귀 모형**

```
Call:
glm(formula = knh20_hchol$HE_HCHOL ~ knh20_hchol$age + as.factor(knh20_hchol$sex) +
    knh20_hchol$BMI + knh20_hchol$HE_wc + as.factor(knh20_hchol$smoke) +
    as.factor(knh20_hchol$BD1), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6090  -0.9037  -0.7266   1.2717   2.1545

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -6.434255   0.429033 -14.997  < 2e-16 ***
knh20_hchol$age               0.025739   0.003414   7.540 4.70e-14 ***
as.factor(knh20_hchol$sex)2   0.840561   0.113506   7.405 1.31e-13 ***
knh20_hchol$BMI              -0.002661   0.023261  -0.114 0.908931
knh20_hchol$HE_wc             0.040933   0.008615   4.752 2.02e-06 ***
as.factor(knh20_hchol$smoke)2 0.261135   0.116220   2.247 0.024646 *
as.factor(knh20_hchol$smoke)3 0.467274   0.127625   3.661 0.000251 ***
as.factor(knh20_hchol$BD1)2   0.114334   0.103317   1.107 0.268451
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5099.3  on 4005  degrees of freedom
Residual deviance: 4867.0  on 3998  degrees of freedom
AIC: 4883
```

```
Number of Fisher Scoring iterations: 4
```

- **오즈비:**

```
(Intercept)  knh20_hchol$age  as.factor(knh20_hchol$sex)2  knh20_hchol$BMI  knh20_hchol$HE_wc
 0.001605604     1.026072836             2.317667925          0.997342785       1.041782556

as.factor(knh20_hchol$smoke)2  as.factor(knh20_hchol$smoke)3  as.factor(knh20_hchol$BD1)2
                1.298403019                    1.595638129                 1.121126861
```

- **오즈비 신뢰구간:**

```
                                  2.5 %       97.5 %
(Intercept)                  0.0006882322 0.003700742
knh20_hchol$age              1.0192507460 1.032983895
as.factor(knh20_hchol$sex)2  1.8578224099 2.899414468
knh20_hchol$BMI              0.9528489020 1.043843459
knh20_hchol$HE_wc            1.0243798229 1.059571721
as.factor(knh20_hchol$smoke)2 1.0343748918 1.631600002
as.factor(knh20_hchol$smoke)3 1.2427061871 2.049963460
as.factor(knh20_hchol$BD1)2  0.9164217522 1.374174890
```

5-3. 위의 두 모형을 Likelihood ratio test(귀무가설 제시) 및 AIC로 비교하여 더 나은 모형을 선택하시오.

- **Script:**

```
library(lmtest)
lrtest(model_hchol,model_hchol_1)
AIC(model_hchol)
AIC(model_hchol_1)
```

- **설명:**

우선 model_hchol과 model_hchol_1 모형에 대해 LRT를 진행하였다. 이때 귀무가설은 "두 모형의 적합도가 유사하다 (간명한 모형을 선택한다)"이며, 검정 결과 귀무가설이 기각되었다. 이를 통해 model_hchol_1의 설명력이 더 좋다는 판단을 내릴 수 있다.

다음으로 AIC를 구하였고 그 결과 model_hchol의 AIC는 5016.12, model_hchol_1의 AIC는 4883.026이 도출되었다. 이때 후자의 AIC가 더 작고 둘의 차이가 2보다 크기 때문에 model_hchol_1이 유의미하게 더욱 설명력이 좋다고 판단된다.

- Likelihood ratio test:

```
Model 1: knh20_hchol$HE_HCHOL ~ knh20_hchol$age
Model 2: knh20_hchol$HE_HCHOL ~ knh20_hchol$age + as.factor(knh20_hchol$sex) +
knh20_hchol$BMI + knh20_hchol$HE_wc + as.factor(knh20_hchol$smoke) +
as.factor(knh20_hchol$BD1)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -2506.1
2   8 -2433.5  6 145.09  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **AIC:**

```
> AIC(model_hchol)
[1] 5016.12
> AIC(model_hchol_1)
[1] 4883.026
```

6. 위의 풀이 과정에서 작성한 R program(script)을 복사하여 붙여 제출하시오.

```
##1-1
knh20_1=subset(knh20,age>=40)


##1-2
library(prettyR)
freq(knh20_1$BD1)
knh20_1$BD1=as.factor(ifelse(knh20_1$BD1<8,knh20_1$BD1,NA))
freq(knh20_1$BD1)


##1-3
freq(knh20_1$BS3_1)
knh20_1$smoke=ifelse(knh20_1$BS3_1<=2,3,ifelse(knh20_1$BS3_1==3,2,ifelse(knh20_1$BS3_1==8,
1,ifelse(knh20_1$BS3_1<=8,knh20_1$BS3_1,NA))))
freq(knh20_1$smoke)


##1-4
knh20_2=subset(knh20_1,select=c(age,sex,HE_chol,HE_wc,smoke,BD1,HE_ht,HE_wt,HE_HCHOL))
library(psych)
describe(knh20_2$age)
describe(knh20_2$HE_chol)
describe(knh20_2$HE_wc)
describe(knh20_2$HE_ht)
describe(knh20_2$HE_wt)
library(prettyR)
freq(knh20_2$sex)
freq(knh20_2$smoke)
freq(knh20_2$BD1)
freq(knh20_2$HE_HCHOL)
```

```
##2-1
knh20_1$AGEGP=as.factor(ifelse(knh20_1$age<50,'1',ifelse(knh20_1$age<60,'2',ifelse(knh20_1$age<
70,'3',ifelse(knh20_1$age,'4'))))
freq(knh20_1$AGEGP)


##2-2
shapiro.test(knh20_1$HE_chol[knh20_1$AGEGP==1])
shapiro.test(knh20_1$HE_chol[knh20_1$AGEGP==2])
shapiro.test(knh20_1$HE_chol[knh20_1$AGEGP==3])
shapiro.test(knh20_1$HE_chol[knh20_1$AGEGP==4])
boxplot(HE_chol~AGEGP,data=knh20_1)
bartlett.test(HE_chol~AGEGP,data=knh20_1)


##2-3
knh20_HE_chol_anova=aov(HE_chol~as.factor(AGEGP),data=knh20_1)
summary(knh20_HE_chol_anova)


##2-4
TukeyHSD(knh20_HE_chol_anova,'as.factor(AGEGP)',conf.level=0.95)
plot(TukeyHSD(knh20_HE_chol_anova,'as.factor(AGEGP)',conf.level=0.95))



##3-1
knh20_40=subset(knh20_1,knh20_1$AGEGP==1)
shapiro.test(knh20_40$HE_chol)
shapiro.test(knh20_40$HE_wc)
```

```
##3-2
plot(HE_chol~HE_wc,data=knh20_40,xlab='총콜레스테롤',ylab="허리둘레",main='scatter plot between chol and wc',type='p',pch=20,cex=1)


##3-3
cor(knh20_40[c('HE_chol','HE_wc')],use='complete.obs',method='pearson')
cor(knh20_40[c('HE_chol','HE_wc')],use='complete.obs',method='spearman')
cor.test(~HE_chol+HE_wc,data=knh20_40,method=c('pearson'))
cor.test(~HE_chol+HE_wc,data=knh20_40,method=c('spearman'))


##4-1
knh20_chol=na.omit(subset(knh20_1,select=c('AGEGP','age','sex','HE_wt','HE_ht','HE_wc','smoke','BD1','HE_chol')))
knh20_chol$BMI=(knh20_chol$HE_wt)/((knh20_chol$HE_ht/100)^2)
describe(knh20_chol$BMI)


##4-2.
library(Hmisc)
knh20_matrix=as.matrix(knh20_chol[c('HE_chol','age','sex','BMI','HE_wc','smoke','BD1')])
rcorr(knh20_matrix,type='pearson')
rcorr(knh20_matrix,type='spearman')
knh20_chol$smoke=as.numeric(knh20_chol$smoke)
knh20_chol$BD1=as.numeric(knh20_chol$BD1)
library(corrplot)
knh20_cor_p=cor(knh20_chol[c('HE_chol','age','sex','BMI','HE_wc','smoke','BD1')],use='complete.obs',method=c('pearson'))
knh20_cor_p
knh20_cor_s=cor(knh20_chol[c('HE_chol','age','sex','BMI','HE_wc','smoke','BD1')],use='complete.obs',method=c('spearman'))
```

```
knh20_cor_s
plot(knh20_chol[c('HE_chol','age','sex','BMI','HE_wc','smoke','BD1')],pch=8,cex=0.5,ellipse=T)


###4-3-1
model_chol=lm(HE_chol~AGEGP,data=knh20_chol)
print(model_chol)
summary(model_chol)
anova(model_chol)


###4-3-2
summary(model_chol)


###4-3-3
par(mfrow=c(2,2))
plot(model_chol)
resid_chol=resid(model_chol)
stdres_chol=rstandard(model_chol)
pred_chol=predict(model_chol)
par(mfrow=c(1,1))
hist(stdres_chol)
library(ggplot2)
ggplot(knh20_chol)+geom_point(mapping=aes(x=AGEGP,y=resid_chol))+geom_hline(yintercept=0,
linetype='dashed', color='red', size=0.5)

##4-4
model_chol_1=lm(HE_chol~AGEGP+factor(sex)+BMI+HE_wc+factor(smoke)+factor(BD1),data=knh2
0_chol)
print(model_chol_1)


###4-4-1
```

```
install.packages('olsrr')
library(olsrr)
ols_vif_tol(model_chol_1)


###4-4-2
anova(model_chol_1)
summary(model_chol_1)


###4-4-3
summary(model_chol_1)


###4-4-4
library(lm.beta)
model_chol.beta=lm.beta(model_chol_1)
print(model_chol.beta)


###4-5-1
summary(model_chol)
summary(model_chol_1)


###4-5-2
anova(model_chol,model_chol_1)



##5-1.
knh20_hchol=na.omit(subset(knh20_1,select=c('HE_HCHOL','age','sex','HE_wt','HE_ht','HE_wc','smoke'
,'BD1')))
knh20_hchol$BMI=(knh20_hchol$HE_wt)/((knh20_hchol$HE_ht/100)^2)
```

```
model_hchol=glm(knh20_hchol$HE_HCHOL~knh20_hchol$age,family=binomial)
summary(model_hchol)
exp(coef(model_hchol))
exp(confint(model_hchol))


##5-2.
model_hchol_1=glm(knh20_hchol$HE_HCHOL~knh20_hchol$age+as.factor(knh20_hchol$sex)+knh2
0_hchol$BMI+knh20_hchol$HE_wc+as.factor(knh20_hchol$smoke)+as.factor(knh20_hchol$BD1),fami
ly=binomial)
summary(model_hchol_1)
exp(coef(model_hchol_1))
exp(confint(model_hchol_1))


##5-3.
library(lmtest)
lrtest(model_hchol,model_hchol_1)
AIC(model_hchol)
AIC(model_hchol_1)
```