지역사회 건강조사자료를 활용한
관심 가설 발표 PPT

20182588 컴퓨터학부 김민수

발표 목차

- 관심 분야 및 관심가설
- 데이터 출처
- 데이터 정제 및 전처리
- 데이터 검정 과정 및 분석 결과
  - 모수/비모수 검정
  - 상관/회귀 분석
  - 다중 회귀 분석
  - 범주형 자료 분석

관심 분야

- 관심 대상 : 만 19세 이상의 성인

- 주된 관심사1 : 비만의 원인
- 주된 관심사2 : 흡연 여부 및 흡연량

관심 가설

- 귀무 가설(H_0)
  - 성인의 흡연 유무 및 흡연량과 비만율은 상관이 없다.

- 대립 가설(H_1)
  - 성인의 흡연 유무 및 흡연량과 비만율은 상관이 있다.

Data 출처

- 통계 분석 예정 Dataset URL :
  https://chs.kdca.go.kr/chs/
- 출처 : KDCA 제공
- DATA : 지역사회건강조사 자료

데이터 정제 및 전처리

- 가독성 높은 칼럼명으로 재 지정
- 키, 몸무게 칼럼의 이상치 제거
- 흡연량 칼럼 답변 수정 및 이상치 제거
  - 비흡연자의 흡연량 문항 답변을 888(비해당) 에서 0개비로 수정
- 흡연 여부 칼럼 이상치 제거
- 금연군이 포함된 흡연 그룹 칼럼 생성
- 키와 몸무게 칼럼을 사용해 BMI 칼럼 생성
- 비만 칼럼 생성
- 성별, 연령대 그룹 칼럼 생성

-1-1	*smb_01z1*	일반담배(궐련) 매일흡연자의 하루 평균 흡연량	□□□개비 777. 응답거부, 888. 비해당(문항1/③, 문항1-1/②③), 999. 모름	N	3

Parametric methods

- 2 groups
  - (0 : 비흡연 그룹, 1 : 흡연 그룹)

```
> var.test(BMI~smokeGroup2, data=data2, conf.level = 0.95 )

        F test to compare two variances

data:  BMI by smokeGroup2
F = 0.93807, num df = 188557, denom df = 36692, p-value = 1.441e-15
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9233071 0.9529797
sample estimates:
ratio of variances
         0.9380662
```

```
> t.test(BMI~smokeGroup2, data=data2, var.equal=F, conf.level = 0.95)

        Welch Two Sample t-test

data:  BMI by smokeGroup2
t = -31.813, df = 50980, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6626911 -0.5857735
sample estimates:
mean in group 0 mean in group 1
       23.52366        24.14790
```

- 3 groups
  - (0 : 비흡연 그룹, 1 : 흡연 그룹, 2 : 금연 그룹)

```
> bartlett.test(BMI~smokeGroup, data=data2)

        Bartlett test of homogeneity of variances

data:  BMI by smokeGroup
Bartlett's K-squared = 525.52, df = 2, p-value < 2.2e-16
```

```
> TukeyHSD(out)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = BMI ~ factor(smokeGroup), data = data2)

$`factor(smokeGroup)`
         diff        lwr       upr p adj
1-0 0.8479574 0.80216498 0.8937497 0e+00
2-0 0.9785923 0.93560677 1.0215778 0e+00
2-1 0.1306349 0.07495878 0.1863111 1e-07
```

표본개수 : 225251
샘플의 크기가 충분히 크기 때문에 정규성 분포는 생략하였고 모분산 검정 및 평균값 검정 결과
p-value < 0.05로 유의미한 차이가 있다.

Non-parametric methods

- 2 groups
  - (0 : 비흡연 그룹, 1 : 흡연 그룹)

```
        Wilcoxon rank sum test with continuity correction

data:  BMI by smokeGroup2
W = 3090734508, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

P-value < 0.05 이므로
두 그룹 간의 중앙값은 유의미한 차이가 난다는 것을 알 수 있습니다.

그룹별 p-value가 0.05보다 작으므로
3그룹 다 중앙값에 유의미한 차이가 난다는 것을 알 수 있습니다.

- 3 groups
  - (0 : 비흡연 그룹, 1 : 흡연 그룹, 2 : 금연 그룹)

```
> summary(result)

 #----------------Nonparametric Multiple Comparisons for relative effects--------------#

 - Alternative Hypothesis:  True differences of relative effects are not equal to 0
 - Estimation Method: Global Pseudo ranks
 - Type of Contrast : Tukey
 - Confidence Level: 95 %
 - Method = Fisher with 57774 DF

 #-------------------------------------------------------------------------------------#

 #----Data Info--------------------------------------------------------------------------#
  Sample   Size    Effect     Lower     Upper
1      0 145450 0.4421396 0.4407837 0.4434963
2      1  36693 0.5181729 0.5163112 0.5200342
3      2  43108 0.5396875 0.5379760 0.5413980

 #----Contrast--------------------------------------------------------------------------#
      1  2 3
2 - 1 -1  1 0
3 - 1 -1  0 1
3 - 2  0 -1 1

 #----Analysis--------------------------------------------------------------------------#
     Estimator Lower Upper Statistic p.Value
2 - 1     0.076 0.072 0.080    44.947       0
3 - 1     0.098 0.094 0.101    64.717       0
3 - 2     0.022 0.017 0.026    10.693       0

 #----Overall---------------------------------------------------------------------------#
  Quantile p.Value
1 2.333271       0

 #-------------------------------------------------------------------------------------#
```

Parametric methods

- 더미 테이블

그룹	N	BMI 평균(±표준편차)	P-value
비흡연	188558	23.52366±3.367624	<2.2e-16
흡연	36693	24.14790±3.456368	

그룹	N	BMI 평균(±표준편차)	P-value
비흡연	145450	23.29994±3.380885	0e+00
흡연	36693	24.14790±3.456368	0e+00
금연	43108	24.27853±3.11657	1e-07

Non-parametric methods

- 더미 테이블

그룹	N	BMI Median(min~max)	P-value
비흡연	188558	23.29123(11.34~65.76)	<2.2e-16
흡연	36693	23.83673(10.02~67.19)	

그룹	N	BMI Median(min~max)	P-value
비흡연	145450	23.01118(11.34~65.76)	0
흡연	36693	23.83673(10.02~67.19)	0
금연	43108	24.09297(11.72~47.88)	0

Correlation

- Pearson 선형 검정 결과

```
        Pearson's product-moment correlation

data:  data3$BMI and data3$smokeAmount
t = 35.284, df = 221667, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07059121 0.07887052
sample estimates:
       cor
0.07473215
```

R = 0.07473215, p-value < 2.2e-16
두 칼럼간 상관관계는 존재하며
양의 상관관계를 보여준다.

- Spearman 선형 검정 결과

```
        Spearman's rank correlation rho

data:  BMI and smokeAmount
S = 1.6884e+15, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06994242
```

R = 0.06994242, p-value < 2.2e-16
두 칼럼간 상관관계는 존재하며
양의 상관관계를 보여준다.

Regression

- dep. var : continuous
- indep. var : continuous

```
Call:
lm(formula = BMI ~ smokeAmount, data = data3)

Residuals:
    Min      1Q  Median      3Q     Max
-14.343  -2.274  -0.274   1.926  42.238

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.526394   0.007612 3090.61   <2e-16 ***
smokeAmount  0.042023   0.001191   35.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.362 on 221667 degrees of freedom
Multiple R-squared:  0.005585,  Adjusted R-squared:  0.00558
F-statistic:  1245 on 1 and 221667 DF,  p-value: < 2.2e-16
```

추정식은 다음과 같으며
BMI = 23.526394 + 0.042023 * SmokeAmount
흡연량이 1개비 증가할 때마다 BMI는 0.042023 증가한다.

Regression

- dep. var : continuous
- indep. var : categorical (2 groups)

```
Call:
lm(formula = BMI ~ smokeGroup2, data = data4)

Residuals:
    Min      1Q  Median      3Q     Max
-14.124  -2.273  -0.273   1.914  43.039

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.523663   0.007751 3035.06   <2e-16 ***
smokeGroup2  0.624232   0.019203   32.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.366 on 225249 degrees of freedom
Multiple R-squared:  0.004669,  Adjusted R-squared:  0.004665
F-statistic:  1057 on 1 and 225249 DF,  p-value: < 2.2e-16
```

추정식은 다음과 같으며,
BMI = 23.534808 + 0.624232 * smokeGroup2
비흡연자 : BMI = 23.534808
흡연자 : BMI = 24.15904

- dep. var : continuous
- indep. var : categorical (>= 3 groups)

```
Call:
lm(formula = BMI ~ factor(smokeGroup), data = data5)

Residuals:
    Min      1Q  Median      3Q     Max
-14.124  -2.240  -0.272   1.895  43.039

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         23.29994    0.00877 2656.92   <2e-16 ***
factor(smokeGroup)1  0.84796    0.01954   43.40   <2e-16 ***
factor(smokeGroup)2  0.97859    0.01834   53.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.345 on 225248 degrees of freedom
Multiple R-squared:  0.01709,   Adjusted R-squared:  0.01708
F-statistic:  1958 on 2 and 225248 DF,  p-value: < 2.2e-16
```

비흡연자 : BMI = 23.29994
흡연자 : BMI = 24.1479
금연 : BMI = 24.27853

Regression

- Exsmoker 그룹 추가 분석
  - 참고 논문 1 : https://ir.ymlib.yonsei.ac.kr/handle/22282913/176990
  - 참고 논문 2 : https://www.jstor.org/stable/29509427#metadata_info_tab_contents

- 논문 1에서는 금연군에서의 체중증가는 운동량 증가 없이 음식물 섭취의 증가한 것에 기인한 것으로 본다.
- 논문 2에서는 금연자의 체중증가 이유를 폐 기능 호전에 따른 것으로 보는 견해도 존재한다. (Nemery, 1983)
- 여기서 BMI와 폐 기능은 흡연과 높은 상관관계를 보여주면서 흡연으로 인한 폐 손상은 흡연자에게 체중감소를 유발한다는 것이다.

Table 4. 금연군의 금연 6개월 후의 체중증가량

항 목	표본수(명)	체중증가량	p값
금연 6개월 후의 체중증가	61	2.14㎏±3.3	0.0001
금연전과 현재 체중간 차이	61	3.89㎏±4.6	0.0001

Table 5. 로지스틱분석에 의한 금연 6개월 후 체중증가에 영향을 주는 변수별 비차비

위험요인	회귀계수	표준오차	Odds Ratio	p-value
음식섭취량3	2.191	0.607	8.944*	0.0003
2 LOG L			15.146	
model significance			0.0001	

*p<0.01 가변수처리 : 음식섭취량 3 증가, 매우증가.

Regression

- 더미 테이블

그룹	N	BMI	P-value
비흡연	188558	23.523663	< 2e-16
흡연	36693	+ 0.624232	< 2e-16

그룹	N	BMI	P-value
비흡연	145450	23.29994	< 2e-16
흡연	36693	+ 0.84796	< 2e-16
금연	43108	+ 0.97859	< 2e-16

Multiple Regression

- Continuous var :
  - smokeAmount(흡연량),
  - age(나이)
- Categorical var :
  - sex(성별)
- BMI = 24.6 - 0.00024*smokeAmount - 0.00485*age
  - 여성(sex=2)인 경우 : -1.35560

```
> summary(res)

Call:
lm(formula = BMI ~ smokeAmount + age + factor(sex), data = data3)

Residuals:
    Min      1Q  Median      3Q     Max
-14.394  -2.224  -0.300   1.854  42.746

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.6218366  0.0249379 987.327   <2e-16 ***
smokeAmount   0.0002418  0.0012570   0.192    0.847
age          -0.0048517  0.0004002 -12.122   <2e-16 ***
factor(sex)2 -1.3556060  0.0151010 -89.770   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.301 on 221665 degrees of freedom
Multiple R-squared:  0.04123,   Adjusted R-squared:  0.04122
F-statistic:  3177 on 3 and 221665 DF,  p-value: < 2.2e-16
```

- AIC 분석 결과 :
  - AIC가 낮을 수록 설명력이 좋음
  - 사용 변수가 age + factor(sex) 일 때, 가장 적합함이 나타난다.
  - 또한, 성별 변수가 BMI를 결정하는데 가장 큰 설명력을 가지며, 흡연량은 가장 작은 설명력을 갖는다.
- 따라서, 성별로 나누어 분석하는 것이 효율적

```
Start:  AIC=529504
BMI ~ smokeAmount + age + factor(sex)

              Df Sum of Sq     RSS    AIC
- smokeAmount  1         0 2415989 529502
<none>                     2415989 529504
- age          1      1602 2417591 529649
- factor(sex)  1     87833 2503822 537418

Step:  AIC=529502.1
BMI ~ age + factor(sex)

              Df Sum of Sq     RSS    AIC
<none>                     2415989 529502
- age          1      1612 2417602 529648
- factor(sex)  1    101015 2517004 538580

Call:
lm(formula = BMI ~ age + factor(sex), data = data3)

Coefficients:
  (Intercept)            age   factor(sex)2
    24.623221      -0.004857      -1.356650
```

Multiple Regression

- AIC 분석 결과 :
  - 각각의 그룹으로 나누어 분석하였을 때의 AIC가 더 낮아 좋은 설명력을 보여준다.

```
Start:  AIC=289733.3
BMI ~ smokeAmount + age

              Df Sum of Sq     RSS    AIC
<none>                     1331776 289733
- smokeAmount  1       132 1331908 289743
- age          1     17467 1349243 291303

Call:
lm(formula = BMI ~ smokeAmount + age, data = femaleData)

Coefficients:
(Intercept)  smokeAmount          age
    21.8042       0.0184       0.0215
```

```
Start:  AIC=234097.3
BMI ~ smokeAmount + age

              Df Sum of Sq     RSS    AIC
<none>                     1025178 234097
- smokeAmount  1       279 1025456 234123
- age          1     43148 1068325 238260

Call:
lm(formula = BMI ~ smokeAmount + age, data = maleData)

Coefficients:
(Intercept)  smokeAmount          age
  26.425975    -0.006555    -0.037786
```

Multiple Regression

- Continuous var : smokeAmount(흡연량), age(나이)

- BMI = 21.8 + 0.0184*smokeAmount + 0.0215*age
- 여성

```
Call:
lm(formula = BMI ~ smokeAmount + age, data = femaleData)

Residuals:
    Min      1Q  Median      3Q     Max
-12.203  -2.264  -0.365   1.834  42.864

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.180e+01  3.147e-02 692.853  < 2e-16 ***
smokeAmount 1.840e-02  5.322e-03   3.458 0.000545 ***
age         2.150e-02  5.404e-04  39.779  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.322 on 120648 degrees of freedom
Multiple R-squared:  0.01297,   Adjusted R-squared:  0.01295
F-statistic: 792.5 on 2 and 120648 DF,  p-value: < 2.2e-16
```

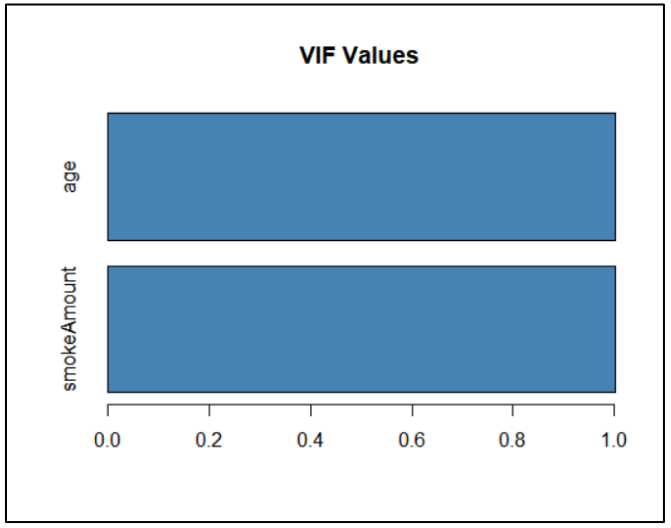
- BMI = 26.42 - 0.0065*smokeAmount – 0.03778*age
- 남성

```
Call:
lm(formula = BMI ~ smokeAmount + age, data = maleData)

Residuals:
    Min      1Q  Median      3Q     Max
-14.646  -2.074  -0.195   1.809  42.082

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.4259751  0.0337483  783.03  < 2e-16 ***
smokeAmount -0.0065554  0.0012510   -5.24 1.61e-07 ***
age         -0.0377857  0.0005795  -65.20  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.186 on 101015 degrees of freedom
Multiple R-squared:  0.04039,   Adjusted R-squared:  0.04037
F-statistic:  2126 on 2 and 101015 DF,  p-value: < 2.2e-16
```

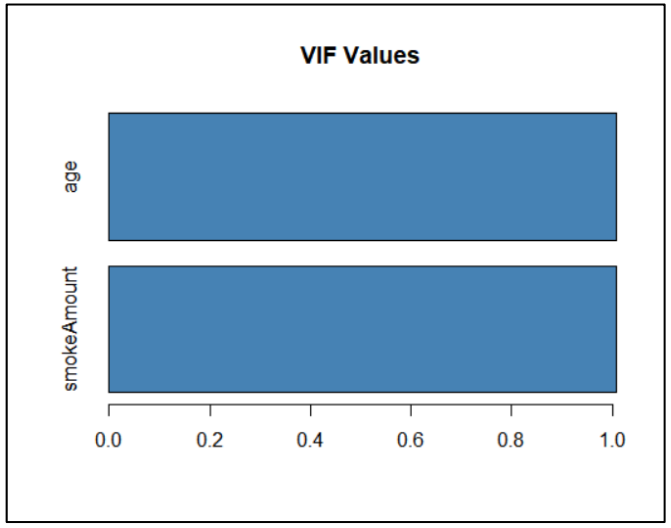
Multiple Regression

- 분산확대인자(VIF)
  - 변수 중 공선성은 존재하지 않음 (VIF가 10을 넘지 않음)

```
smokeAmount          age
   1.002173     1.002173
```




```
smokeAmount          age
   1.008258     1.008258
```



Multiple Regression

- 더미 테이블

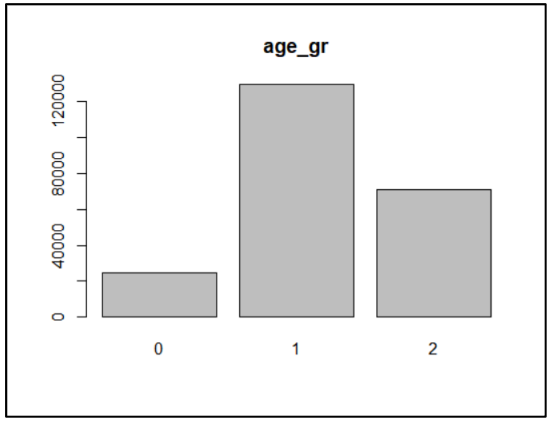
그룹	N	BMI	P-value
남성	101018	BMI = 26.42 - 0.0065*smokeAmount – 0.03778*age	< 2e-16
여성	120651	BMI = 21.8 + 0.0184*smokeAmount + 0.0215*age	< 2e-16

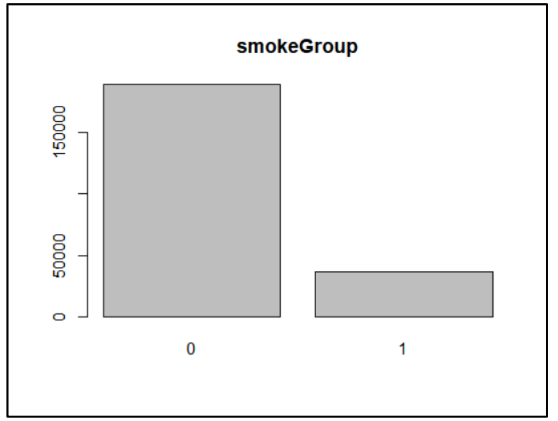
Multiple Regression

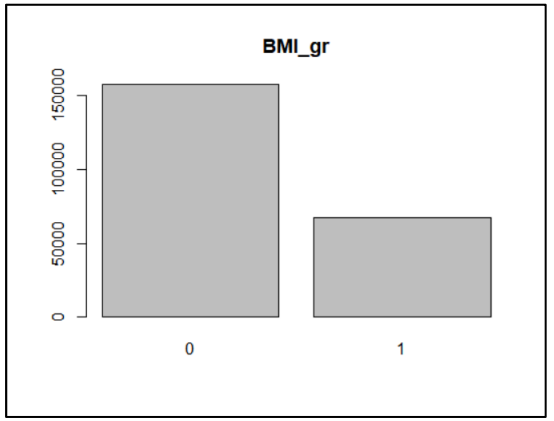
- 성별로 나누지 않고 분석한 경우, AIC가 약 53만 정도로 매우 높아 설명력이 부족하다. 또한 smokeAmount가 다른 변수에 비해 작은 설명력은 갖는다.
- 따라서, 성별로 나눠 분석을 추가적으로 진행
  - 남성/여성 두 그룹 모두 AIC가 많아 낮아져서 상대적으로 높은 설명력을 갖는다.
  - 남성 그룹의 경우, smokeAmount와 age가 음의 기울기를 갖는다.
    - 따라서 남성의 경우, 흡연할 수록 나이가 들수록 BMI가 감소한다.
  - 여성 그룹의 경우, smokeAmount와 age가 양의 기울기를 갖는다.
    - 따라서 여성의 경우, 흡연할 수록 나이가 들수록 BMI가 증가한다.

Categorical Analysis

- 사용 변수 :
  - Age_gr(연령 그룹)
    - 청년(19~29) : 0
    - 중장년(30~64) : 1
    - 노년(65~) : 2
  - smokeGroup(흡연/비흡연)
    - 비흡연 : 0
    - 흡연 : 1
  - BMI_gr
    - 비만 아닌 그룹 : 0
    - 비만 그룹(BMI>=25) : 1







Categorical Analysis

- 카이제곱 검정 결과 비교
  - 2가지 그룹 모두 흡연그룹 변수와 연령대 그룹에 대한 검정 결과에서
  - p-value가 0.05보다 낮은 모습을 보여준다.

```
> chisq.test(femaleData$BMI_gr, femaleData$smokeGroup2)

        Pearson's Chi-squared test with Yates' continuity correction

data:  femaleData$BMI_gr and femaleData$smokeGroup2
X-squared = 4.6501, df = 1, p-value = 0.03105

> chisq.test(femaleData$BMI_gr, femaleData$age_gr)

        Pearson's Chi-squared test

data:  femaleData$BMI_gr and femaleData$age_gr
X-squared = 741.54, df = 2, p-value < 2.2e-16
```

```
> chisq.test(maleData$BMI_gr, maleData$smokeGroup2)

        Pearson's Chi-squared test with Yates' continuity correction

data:  maleData$BMI_gr and maleData$smokeGroup2
X-squared = 11.927, df = 1, p-value = 0.0005532

> chisq.test(maleData$BMI_gr, maleData$age_gr)

        Pearson's Chi-squared test

data:  maleData$BMI_gr and maleData$age_gr
X-squared = 2539.7, df = 2, p-value < 2.2e-16
```

Categorical Analysis

- GLM (Generalized Linear Model)

```
glm(formula = BMI_gr ~ factor(age_gr) + smokeGroup2, family = binomial,
    data = femaleData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8479  -0.7779  -0.7230  -0.5710   1.9465

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.73142    0.02498 -69.322  < 2e-16 ***
factor(age_gr)1  0.52325    0.02648  19.759  < 2e-16 ***
factor(age_gr)2  0.72574    0.02734  26.543  < 2e-16 ***
smokeGroup2     0.16773    0.04086   4.105 4.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132409  on 121389  degrees of freedom
Residual deviance: 131608  on 121386  degrees of freedom
AIC: 131616

Number of Fisher Scoring iterations: 4
```

여성

```
Call:
glm(formula = BMI_gr ~ factor(age_gr) + smokeGroup2, family = binomial,
    data = maleData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0967  -1.0181  -0.7929   1.2603   1.7063

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.42208    0.01925  -21.92   <2e-16 ***
factor(age_gr)1  0.22923    0.02056   11.15   <2e-16 ***
factor(age_gr)2 -0.57392    0.02287  -25.09   <2e-16 ***
smokeGroup2    -0.19404    0.01417  -13.69   <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: 137584  on 103860  degrees of freedom
Residual deviance: 134779  on 103857  degrees of freedom
AIC: 134787

Number of Fisher Scoring iterations: 4
```

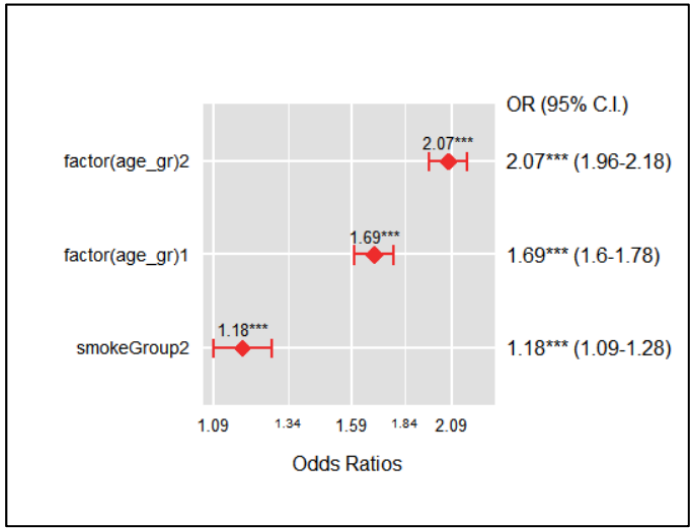
남성

Categorical Analysis

- Odds Ratio table

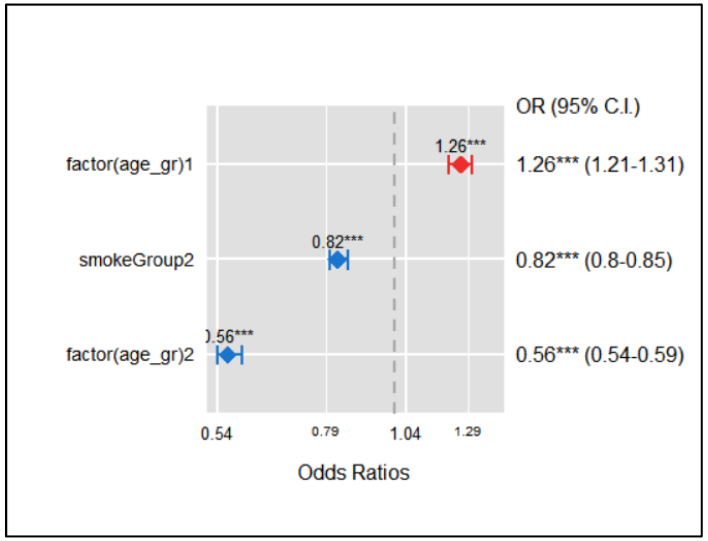
	OR	2.5%	97.5%	p
(Intercept)	0.18	0.17	0.19	0
factor(age_gr)1	1.69	1.60	1.78	0
factor(age_gr)2	2.07	1.96	2.18	0
smokeGroup2	1.18	1.09	1.28	0

여성



	OR	2.5%	97.5%	p
(Intercept)	0.66	0.63	0.68	0
factor(age_gr)1	1.26	1.21	1.31	0
factor(age_gr)2	0.56	0.54	0.59	0
smokeGroup2	0.82	0.80	0.85	0

남성



Categorical Analysis

- Multiple regression과 마찬가지로 성별로 나누지 않고 분석한 경우, 흡연 여부 변수의 설명력이 부족하여 나누어 추가적으로 분석을 진행하였다.
- 남성/여성으로 나눌 경우 smokerGroup 변수는 보다 높은 설명력을 갖는다.

- 여성 그룹의 경우
  - 흡연자는 비흡연자에 비해 비만일 확률이 18% 더 높다.
  - 중장년층은 청년층에 비해 비만일 확률이 69% 더 높다.
  - 노년층은 청년층에 비해 비만일 확률이 107% 더 높다.
- 남성 그룹의 경우
  - 흡연자는 비흡연자에 비해 비만일 확률이 18% 더 낮다.
  - 중장년층은 청년층에 비해 비만일 확률이 26% 더 높다.
  - 노년층은 청년층에 비해 비만일 확률이 44% 더 낮다.