2022-1학기 회귀분석(1139301-01)
Final project
- 국민건강영양조사 자료를 이용하여 -

1인가구의 생활습관에 따른 건강분석

경제학과 20200856 정가연

2022-1학기 회귀분석(1139301-01)
Final project
- 국민건강영양조사 자료를 이용하여 -

1인가구의 생활습관에 따른 건강분석

경제학과 20200856 정가연

목차

1-1. 연구 배경

1인 가구란?
가구원이 한 명인 가구로, 2000년대 이후 증가중

2000년: 15.5%

2005년: 20%

2010년: 23.9%

2015년: 27.2%

현재: 31.7%

1-1. 연구 배경

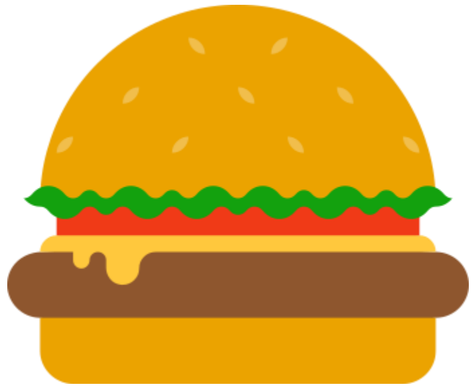
1인 가구의 특징

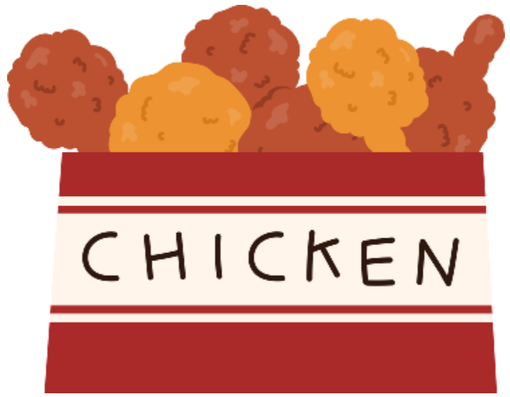
-건강에 관심이 적음

-건강하지 못한 생활 습관

-고나트륨 · 고지방 · 고열량의 식습관

-배달/테이크아웃 이용률 높음





1-2. 연구 목표

[국민건강영양조사 데이터 분석]

1인 가구의 생활습관 조사
+
건강과의 관련성 파악

1인 가구의 건강에 대한 기초 자료 제공

2-1. 연구 내용

연구 도구: 통계프로그램 R의 dplyr, leaps, MASS 패키지

자료: 국민건강영양조사

데이터 전처리 과정: 모름, 무응답 -> 결측치로 제거

데이터 전처리 결과: 최초 표본 8,110명 -> 676명

대상: 1인 가구 676명

2-1. 연구 내용

반응변수: 건강을 나타내는 건강 관련 삶의 질

설명변수: 생활 습관과 관련한 변수들

-> 건강과 생활 습관의 상관관계를 나타내는 회귀모형 제작

-> 생활습관과 관련해서 실제로 1인 가구의 건강이 관련이 있는지 확인 및 어떠한 의미를 가지는지 파악

2-2. 연구 도구

반응변수: EQ5D(건강관련 삶의 질)

표 19. EQ-5D Index (Euro Quality of Life - 5 Dimensions) 변수 생성 - SAS 프로그램

```
/* 1.운동능력 수준 */
IF LQ_1EQL in (1, 2, 3) THEN M2 = (LQ_1EQL=2) ;
IF LQ_1EQL in (1, 2, 3) THEN M3 = (LQ_1EQL=3) ;

/* 2.자기관리 수준 */
IF LQ_2EQL in (1, 2, 3) THEN SC2 = (LQ_2EQL=2) ;
IF LQ_2EQL in (1, 2, 3) THEN SC3 = (LQ_2EQL=3) ;

/* 3.일상활동 수준 */
IF LQ_3EQL in (1, 2, 3) THEN UA2 = (LQ_3EQL=2) ;
IF LQ_3EQL in (1, 2, 3) THEN UA3 = (LQ_3EQL=3) ;

/* 4.통증, 불편감 수준 */
IF LQ_4EQL in (1, 2, 3) THEN PD2 = (LQ_4EQL=2) ;
IF LQ_4EQL in (1, 2, 3) THEN PD3 = (LQ_4EQL=3) ;

/* 5.불안, 우울 수준 */
IF LQ_5EQL in (1, 2, 3) THEN AD2 = (LQ_5EQL=2) ;
IF LQ_5EQL in (1, 2, 3) THEN AD3 = (LQ_5EQL=3) ;

/* Interaction 모형 */
IF LQ_1EQL in (1, 2, 3) & LQ_2EQL in (1, 2, 3) & LQ_3EQL in (1, 2, 3) &
   LQ_4EQL in (1, 2, 3) & LQ_5EQL in (1, 2, 3)
THEN N3 = (LQ_1EQL=3 or LQ_2EQL=3 or LQ_3EQL=3 or LQ_4EQL=3 or LQ_5EQL=3);
```

- Euro-Qol group의 EQ-5D 활용

[5가지 구성]

1.운동능력 수준

2.자기관리 수준

3.일상활동 수준

4.통증, 불편감 수준

5.불안, 우울 수준

[3가지 구성]

1=문제 없음

2=다소 문제 있음

3=심각한 문제 있음

2-2. 연구 도구

반응변수: EQ5D(건강관련 삶의 질)

```
/*>>> EQ-5D index : 삶의 질 조사도구(EQ-5D)의 질 가중치 추정 연구 보고서, 질병관리본부, 2007)*/

EQ5D  =  1 - (0.05 + 0.096*M2 + 0.418*M3 + 0.046*SC2 + 0.136*SC3 + 0.051*UA2 +
              0.208*UA3 + 0.037*PD2 + 0.151*PD3 + 0.043*AD2 + 0.158*AD3 + 0.05*N3) ;

IF LQ_1EQL=1 & LQ_2EQL=1 & LQ_3EQL=1 & LQ_4EQL=1 & LQ_5EQL=1 THEN EQ5D = 1 ;
```

건강상태는 EQ-5D지수의 5가지 영역의 3가지 수준(3^5 =243)으로 나타남

이렇게 얻어진 건강상태에 가중치를 곱해서 계산하며 -1점과 +1점 사이에 분포

EQ5D의 값이 1에 가까울수록 건강상태가 좋음

2-2. 연구 도구

설명변수: 생활습관 관련 변수(10개)

L_OUT_FQ(외식 횟수), BS3_1(현재흡연 여부), BP1(평소 스트레스 인지 정도),
BP16_1((만12세이상)주중(또는 일하는 날)하루 평균 수면 시간),
BP16_2 ((만12세이상) 주말(또는 일하지 않는 날, 일하지 않는 전날) 하루 평균 수면 시간),
LF_S4(식비가 부족하여 균형잡힌 식사를 할 수 없던 경험), DI1_2(혈압조절제 복용),
DI2_2(이상지질혈증 약복용), DJ4_3(천식 약복용(소아, 청소년 포함)),
BE3_31(1주일간 걷기 일수)

2-3. 탐색적 분석

중회귀모형: $Yi = \beta 0 + \beta 1Xi1 + \beta 2Xi2 + \beta 3Xi3 + \beta 4Xi4 + \cdots + + \beta 10Xi10 + \epsilon i,\ i = 1,2, \ldots ,676$

Yi: i번째 조사대상자의 건강관련 삶의 질 지수

$Xi1$: i번째 조사대상자의 외식횟수

$Xi2$: i번째 조사대상자의 현재흡연 여부

$Xi3$: i번째 조사대상자의 평소 스트레스 인지 정도

$Xi4$: i번째 조사대상자의 주중(또는 일하는 날) 하루 평균 수면 시간

$Xi5$: i번째 조사대상자의 주말(또는 일하지 않는 날, 일하지 않는 전날) 하루 평균 수면 시간

$Xi6$: : i번째 조사대상자의 식비가 부족하여 균형잡힌 식사를 할 수 없던 경험

$Xi7$: i번째 조사대상자의 혈압조절제 복용

$Xi8$: i번째 조사대상자의 이상지질혈증 약복용

$Xi9$: i번째 조사대상자의 천식 약복용(소아, 청소년 포함)

$Xi10$: i번째 조사대상자의 1주일간 걷기 일수

2-3. 탐색적 분석- 변수선택 기준

추정회귀계수

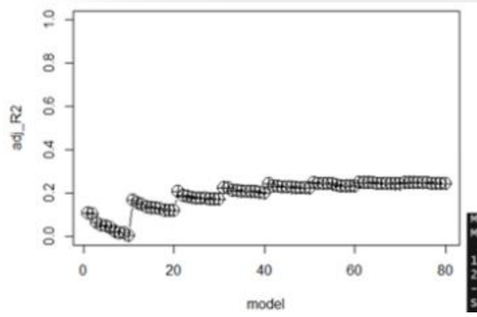
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.599883   0.053901  11.129  < 2e-16 ***
X1          -0.015983   0.003116  -5.128 3.84e-07 ***
X2          -0.004071   0.001734  -2.348  0.01916 *
X3           0.029770   0.006596   4.513 7.54e-06 ***
X4          -0.001433   0.005038  -0.285  0.77611
X5           0.003658   0.004059   0.901  0.36784
X6           0.056306   0.009366   6.012 3.03e-09 ***
X7           0.001797   0.001762   1.020  0.30806
X8           0.006422   0.001967   3.265  0.00115 **
X9           0.004644   0.004389   1.058  0.29048
X10          0.010406   0.001884   5.523 4.77e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1298 on 665 degrees of freedom
Multiple R-squared:  0.2587,    Adjusted R-squared:  0.2476
F-statistic: 23.21 on 10 and 665 DF,  p-value: < 2.2e-16
```

[기준: 유의수준 5%]

유의: L_OUT_FQ(X1), BS3_1(X2), BP1(X3), LF_S4 (X6), DI2_2(X8), BE3_31(X10)

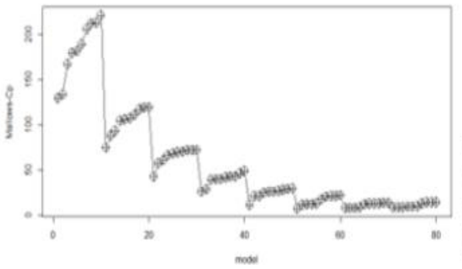
유의X: BP16_1(X4), BP16_2(X5), DI1_2 (X7), DJ4_3(X9)

2-3. 탐색적 분석 - 변수선택 기준



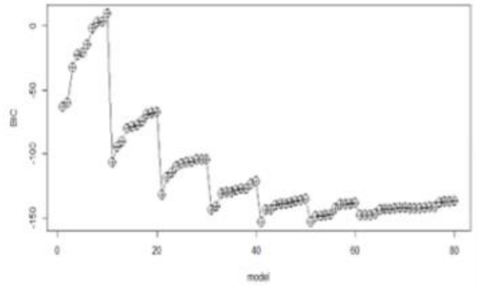

```
Model 1: EQ5D ~ L_OUT_FQ + BS3_1 + BP1 + LF_S4 + DI2_2 + BE3_31
Model 2: EQ5D ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    669 11.259
2    675 15.108 -6   -3.8485 38.111 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

수정된 결정계수 adj_R -> 71번째 모형




```
Model 1: EQ5D ~ L_OUT_FQ + BS3_1 + BP1 + BP16_2 + LF_S4 + DI2_2 + DJ4_3 +
    BE3_31
Model 2: EQ5D ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    667 11.219
2    675 15.108 -8   -3.8894 28.905 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mallows-Cp-> 51번째 모형




```
Model 1: EQ5D ~ L_OUT_FQ + BS3_1 + BP1 + BP16_2 + LF_S4 + DI2_2 + DJ4_3 +
    BE3_31
Model 2: EQ5D ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    667 11.219
2    675 15.108 -8   -3.8894 28.905 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

BIC-> 51번째 모형

적합한 변수: X1, X2, X3, X6, X8, X10

2-3. 탐색적 분석 - 변수선택 방법

단계별 회귀

```
Call:
lm(formula = EQ5D ~ L_OUT_FQ + LF_S4 + BE3_31 + BP1 + DI2_2 +
    BS3_1, data = model)

Coefficients:
(Intercept)     L_OUT_FQ        LF_S4       BE3_31          BP1        DI2_2        BS3_1
   0.662185    -0.017859     0.057101     0.010837     0.029470     0.007410    -0.004323
```

전진선택법

```
Call:
lm(formula = EQ5D ~ L_OUT_FQ + LF_S4 + BE3_31 + BP1 + DI2_2 +
    BS3_1, data = model)

Coefficients:
(Intercept)     L_OUT_FQ        LF_S4       BE3_31          BP1        DI2_2        BS3_1
   0.662185    -0.017859     0.057101     0.010837     0.029470     0.007410    -0.004323
```

후진제거법

```
Call:
lm(formula = EQ5D ~ L_OUT_FQ + BS3_1 + BP1 + LF_S4 + DI2_2 +
    BE3_31, data = model)

Coefficients:
(Intercept)     L_OUT_FQ        BS3_1          BP1        LF_S4        DI2_2       BE3_31
   0.662185    -0.017859    -0.004323     0.029470     0.057101     0.007410     0.010837
```

적합한 변수: X1, X2, X3, X6, X7, X10

2-3. 탐색적 분석 - 회귀모형 진단 및 수정

가설 : $H0 : \beta4 = \beta5 = \beta7 = \beta9 = 0$ VS $Ha : not\ H0$

```
Model 1: Y ~ X1 + X2 + X3 + X6 + X8 + X10
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    669 11.259
2    665 11.199  4  0.060429 0.8971 0.4652
```

[변수선택기준 결과]

-> 귀무가설 기각 X

->Reduced_model 적합

가설 : $H0 : \beta4 = \beta5 = \beta8 = \beta9 = 0$ VS $Ha : not\ H0$

```
Model 1: Y ~ X1 + X2 + X3 + X6 + X7 + X10
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    669 11.435
2    665 11.199  4   0.23603 3.5039 0.007652 **
```

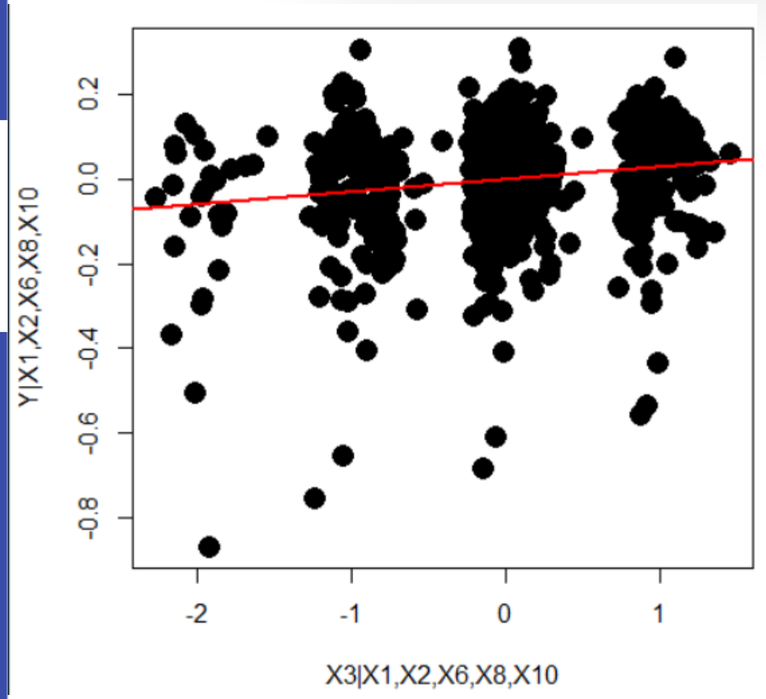
변수선택방법 결과

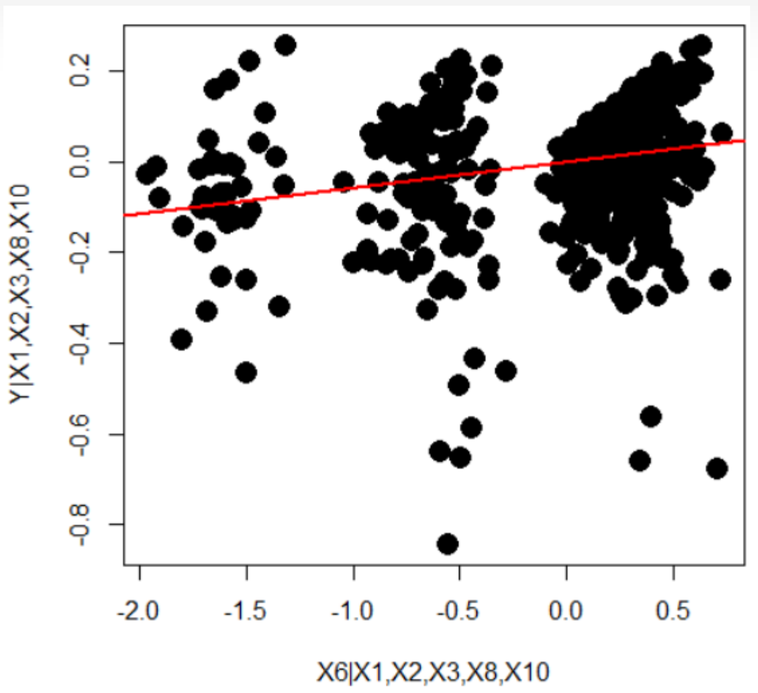
-> 귀무가설 기각 O

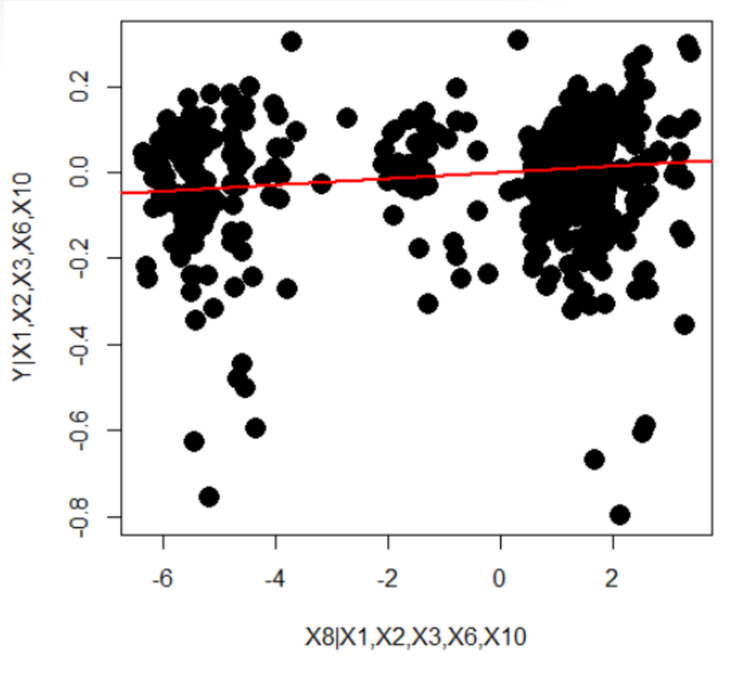
->Full_model 적합

적합한 변수: X1, X2, X3, X6, X8, X10

2-3. 탐색적 분석 - 편회귀그림







가장 뚜렷한 선형관계를 보인 변수: X6

순수한 X6의 값이 반응변수 Y에 가장 많은 영향

2-3. 탐색적 분석 - 적합결여검정

```
Model 1: Y ~ X1
Model 2: Y ~ factor(X1)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    674 13.492
2    669 13.199  5   0.29213 2.9613 0.01183
```

```
Model 1: Y ~ X3
Model 2: Y ~ factor(X3)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    674 14.882
2    672 14.601  2   0.28161 6.4806 0.001631
```

```
Model 1: Y ~ X10
Model 2: Y ~ factor(X10)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    674 14.127
2    668 13.773  6    0.3539 2.8607 0.00929
```

X1, X3, X10의 적합결여검정 결과 귀무가설 기각 O

S^2이 과대추정 O

X1, X3, X10은 적합하지 않은 변수로 판단

2-3. 탐색적 분석 - 적합결여검정

```
Model 1: Y ~ X2
Model 2: Y ~ factor(X2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    674 14.774
2    672 14.765  2 0.0086928 0.1978 0.8206
```

```
Model 1: Y ~ X6
Model 2: Y ~ factor(X6)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    674 13.563
2    673 13.502  1  0.061222 3.0516 0.08112
```

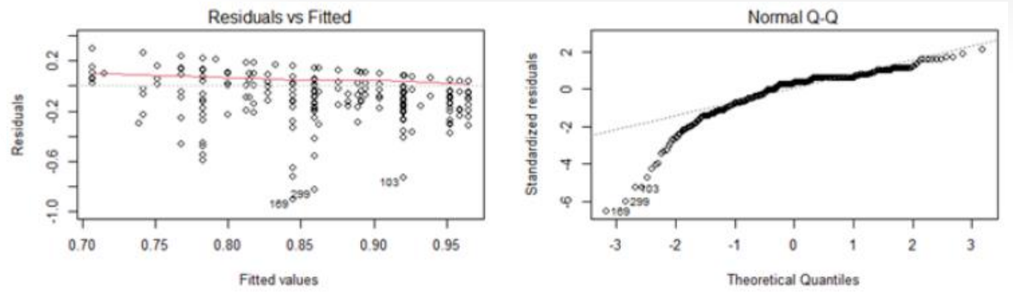
```
Model 1: Y ~ X8
Model 2: Y ~ factor(X8)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    674 14.372
2    671 14.332  3   0.03932 0.6136 0.6063
```

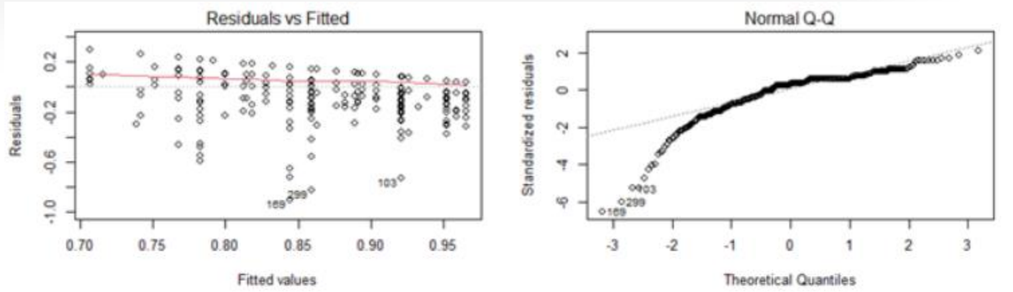
X2, X6, X8의 적합결여검정 결과 귀무가설 기각 X

S^2이 과대추정 X

X2, X6, X8은 적합한 변수로 판단

2-3. 탐색적 분석 - 변수변환





[변수 변환 전]

[√Y 변환 후]

변수변환 후와 큰 차이 없으므로 변수변환 전 모델인

Reduced3_model을 최종모형 후보로 결정

2-4. 회귀 모형 선택 - 모형 검정

[다중공선성 탐색]

다중공선성이란?

설명변수들간의 상관관계가 높아 최소제곱추정량의 계산이 불가능해지는 것을 의미한다.

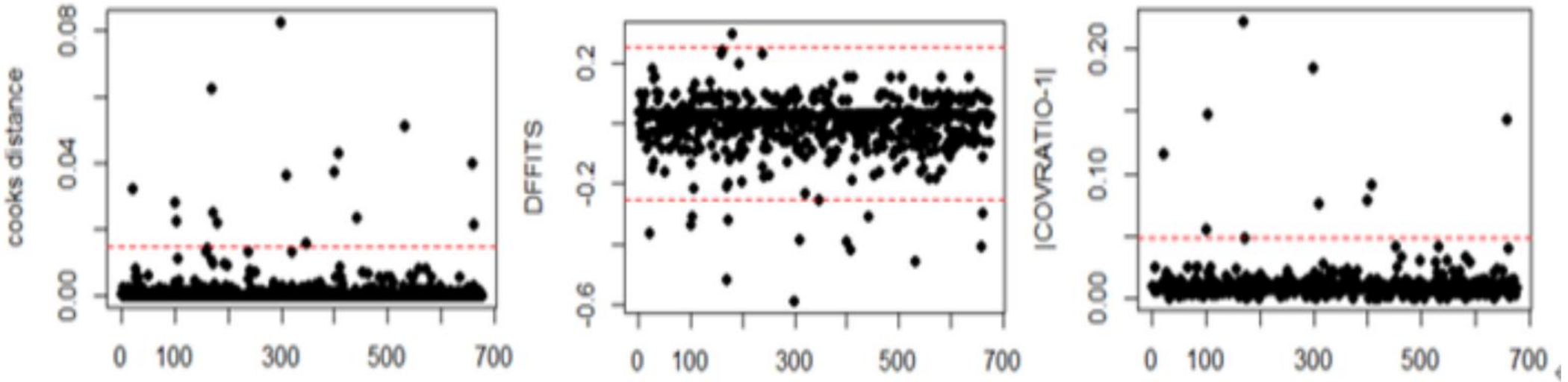
```
> vif(Reduced3_model)
       X2        X6        X8
1.005010 1.033535 1.036958
```

VIF <5이므로

다중공선성 X

2-4. 회귀 모형 선택 - 모형 검정

[영향력 측도]



676개의 행으로 구성되었던 Reduced3_model에서 **63**개의 이상치를 제외하고 **613**개의 행으로 구성된 final_model을 최종모형으로 결정

2-4. 회귀 모형 선택 - 모형 검정

[더빈-왓슨 검정]

```
        Durbin-Watson test

data:  Final_model
DW = 1.9761, p-value = 0.3754
```

더빈-왓슨 검정통계량은 0~4사이에 있으므로 오차항의 독립을 만족한다

2-4. 회귀 모형 선택 - 최종 모형 결정

```
> #PRESS VS SSE
> PRESS_last$stat
[1] 6.090048
> SSE
[1] 5.987214
> #R2 VS R2_predic
> 1-(SSE/SST)
[1] 0.1175041
> 1-(PRESS_last$stat/SST)
[1] 0.1023467
```

훈련자료(70%, 429개)

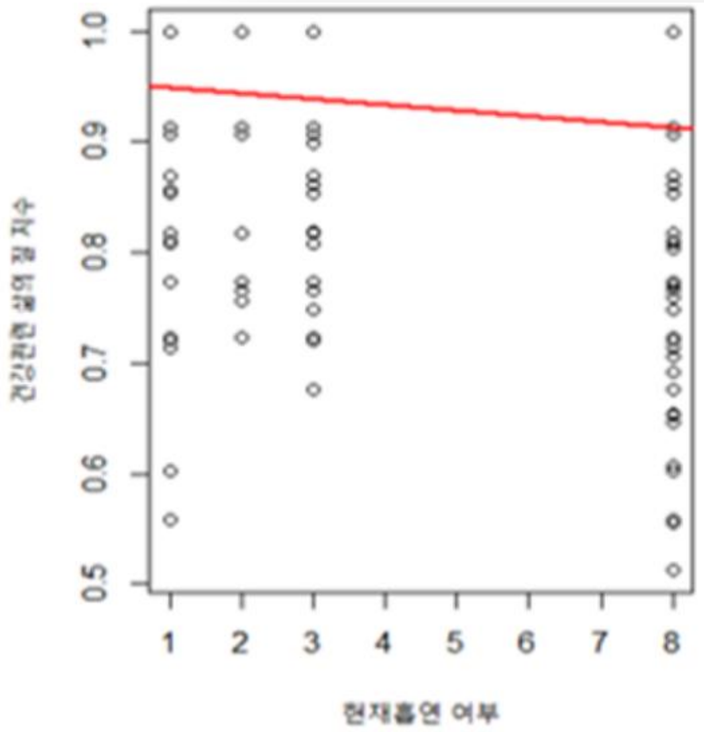
확인자료(30%, 184개)

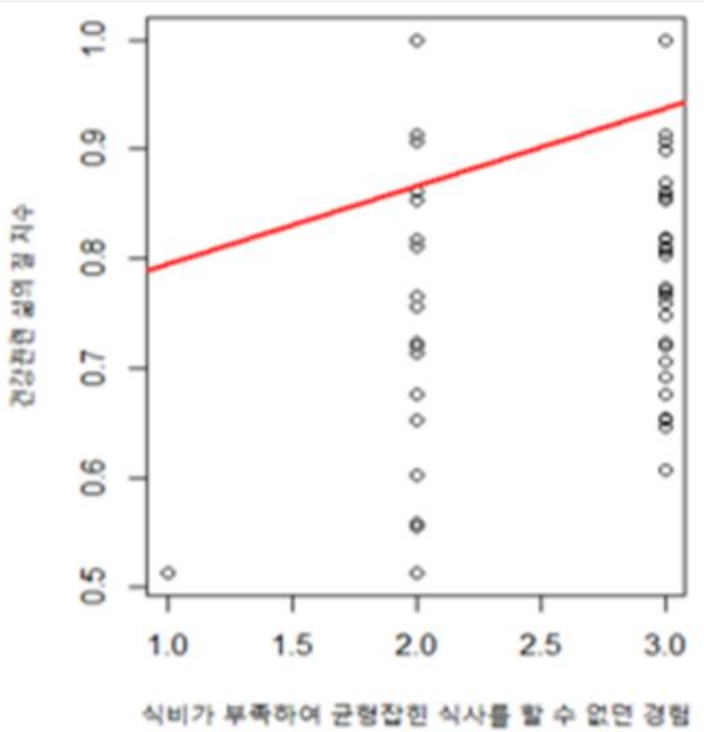
PRESS VS SSE -> 비슷

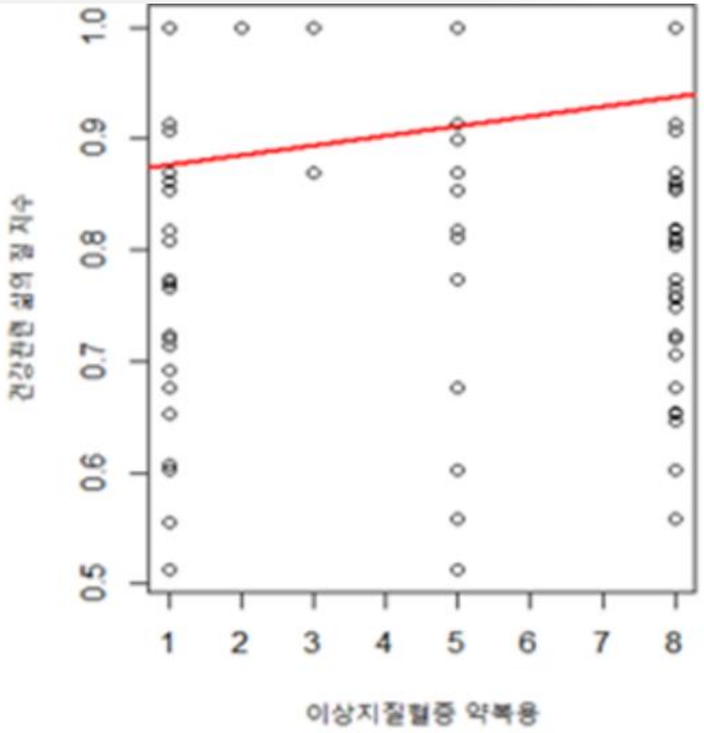
R^2 VS R^2predict -> 비슷

최종모형은 타당하다

2-4. 회귀 모형 선택 - 최종 모형 결정







-흡연 횟수 ⬇

-건강 관련 삶의 질 ⬆

-균형잡힌 식사 경험 ⬇

-건강 관련 삶의 질 ⬆

-이상지질혈증 약복용 횟수 ⬇

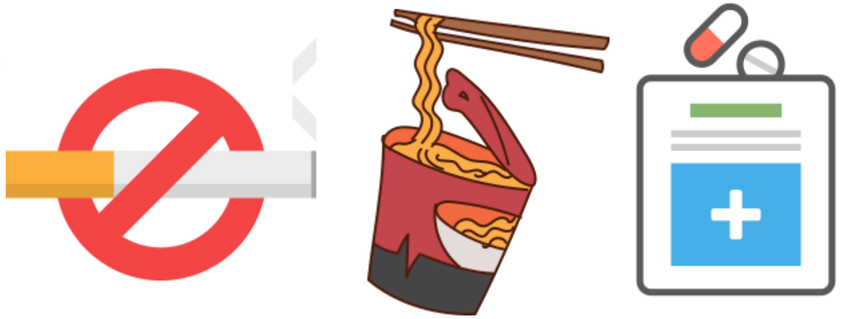
-건강 관련 삶의 질 ⬆

3. 결론 - 내용요약

-1인 가구들의 건강 관련 삶의 질에 영향을 주고 있는 요인들을 파악

-현재 흡연 여부, 식비가 부족하여 균형잡힌 식사를 할 수 없던 경험,
이상 지지혈증 약 복용이 1인 가구의 건강 관련 삶에 영향을 준다는 것을 확인

-1인 가구의 건강에 대한 자료를 제공하였고 1인 가구의 건강 관련 삶의 질을 향상하기 위한 연구가 필요하다고 전하고 싶다.



3. 결론 - 느낀점

1. 선형성, 독립성, 다중공선성 등을 확인하고 회귀진단을 실시하면서 유의미한 변수들이 감소하게 되었고, 이를 보면서 예비모형을 구축할 때 더욱 많은 설명변수를 관측해야 했음.
2. 최종 회귀모형을 확인해보면 X2 변수와 X8 변수에서 8 비해당 부분에 자료가 많이 몰려있었음. 모름, 무응답과 함께 결측치 처리를 해야 했음.

다음 프로젝트를 진행할 때는 이러한 한계점을 보완하며 더 정확하고 유의미한 회귀모형을 만들 것이다.

4. 참고문헌

서울&, “저소득 가구 여성과 1인 가구, 서울시민 중 가장 건강 취약”, 2022-02-24

김종규,권이승,가천대학교 경상대학 헬스케어경영학과, 연세대학교 보건과학대학, “보건행정학 EQ-5DIndex 이용 성인 암 환자의 인구사회학적 특성별건강관련 삶의 질 측정”

KOSIS, “가구주의 성, 연령 및 세대구성별 가구(일반가구)”

경청해주셔서
감사합니다.

경청해주셔서
감사합니다.