1인 가구의 생활습관에 따른 건강 분석

▼ I 서론

요약

국민건강영양조사 자료를 바탕으로 연구를 진행하였다. 원본 자료에서 관심 있는 변수들만 추출한 후 모름과 무응답은 결측지로 모두 제거한 후 회귀분석을 진행하였다. 1인 가구의 생활 습관과 건강과의 상관관계를 확인 후 1인 가구 건강에 대한 기초자료를 제공하자 한다.

1. 연구 배경

1인 가구란 “가구원이 한 명인 가구로, 2000년대 이후 결혼 시기가 늦춰지고 미혼율 및 이혼율 증가와 함께 사회가 고령화되면서 그 비중이 높아지고 있다.” 통계청의 인구총조사에 따르면 “2000년도 15.5%에서 2005년도에 20%로, 2010년에는 23.9%, 2015년 27.2%”로 지속해서 증가하고 있으며 현재는 31.7%로 높은 비율을 차지하고 있다. 1인 가구들은 건강에 관한 관심이 적은 편이며 식사가 규칙적이지 못하고 건강한 생활 습관을 가지지 못한다고 알려져 있다. 또한 외식과 배달 · 테이크아웃의 이용률이 높으므로 고나트륨 · 고지방 · 고열량의 음식을 섭취 비율이 높으며 안 좋은 식습관을 가지곤 한다. 1인 가구들은 본인의 학업이나 직장 때문에 1인 가구가 되는 경우가 많으므로 다른 연령대에 비해서 본인의 건강을 살피는데 여유가 없다고 한다.

현재 필자 또한 기숙사에 거주하며 1인 가구의 형태를 취하고 있는데 건강한 음식을 챙겨 먹는 것은 쉽지 않은 게 현실이다. 생활습관이 나빠지면서 체력이 안좋아지는 것을 느끼게 되었고 다른 1인가구의 생활 습관과 건강에 대해 알아보고자 하는 취지에서 연구 주제를 정하게 되었다.

1. 연구 목표

국민건강영양조사 데이터를 분석함으로써 1인 가구의 생활습관을 알아본 후 이들의 건강과 어떤 관련이 있는지 파악해볼 것이다. 1인 가구 생활 습관과 건강에 대한 분석을 통해서 1인 가구의 건강에 대한 자료를 제공하고자 한다.

▼ II 본론

1. 연구 내용

통계프로그램인 R의 dplyr, leaps, MASS 등을 이용했고 국민건강영양조사 자료를 전처리한 후 회귀모형을 만들어보았다. 1인 가구 형태인 사람을 대상으로 정하였다. 그 모름, 무응답은 결측치로 제거하며 데이터 전처리를 진행해 준 결과 최초 표본이었던 8,110명에서 676명으로 감소하였고 이들을 분석대상자로 정했다.

건강을 나타내는 건강 관련 삶의 질을 반응변수로 두고 생활 습관과 관련한 변수들을 설명변수로 정하였다. 반응변수와 설명변수의 관계, 즉 건강과 생활 습관의 상관관계를 나타내는 회귀모형을 만들어보

았다. 최종으로 완성된 회귀모형을 점검하고 생활습관과 관련해서 실제로 1인 가구의 건강이 관련이 있는지 확인하고 어떠한 의미를 나타내는지 파악하며 결론을 내렸다.

1. **연구 도구**

2-1) 반응 변수 및 설명 변수

반응변수로는 EQ5D(건강관련 삶의 질)을 사용하였다. 이는 Euro-QoI group에서 제작한 EQ-5D를 활용한 것으로, 국민건강영양조사에서 2005년부터 사용하기 시작하였다. EQ5D는 1.운동능력 수준, 2.자기관리 수준, 3.일상활동 수준, 4.통증, 불편감 수준, 5.불안, 우울 수준으로 구성되어 있고 1=문제 없음, 2=다소 문제 있음, 3=심각한 문제 있음의 3가지 수준으로 나타난다.

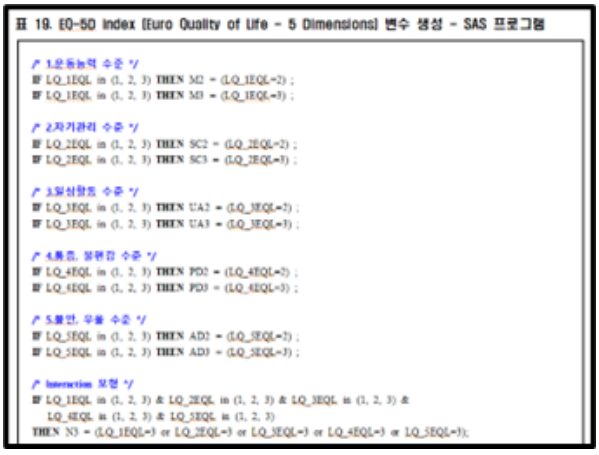


표 19. EQ-5D Index (Euro Quality of Life - 5 Dimensions) 변수 생성 - SAS 프로그램

```
/* 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 포함 */
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);
```

건강관련 삶의 질 지수는 건강상태를 EQ-5D지수의 5가지 영역의 3가지 수준(3^5 =243)으로 나타낸다. 이렇게 얻어진 건강상태에 가중치를 곱해서 계산하며 –1점과 +1점 사이에 분포되어 있다. EQ5D의 값이 1에 가까울수록 건강상태가 좋은 것으로 본다.

```
/*>>> 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 ;
```

설명변수로는 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) 을 사용하였다.

2-2) 분석 방법

통계 프로그램인 R을 이용해서 탐색적 분석을 한 후 회귀모형을 검정하는 순으로 분석을 진행하였다. 변수선택 기준으로 수정된 결정계수, Mallows – Cp, PRESS, AIC, BIC을 사용한 후 예비모형을 구축하였다. 부분F검정, 편회귀그림, 적합결여검정, 변수변환을 통해 회귀모형을 진단하고 수정하였다. 탐색적 분석을 바탕으로 핵심적인 설명변수와 회귀모형을 선정한 후 표준화 회귀계수, 다중공선성 탐색을 통해서 회귀진단을 실시하며 최종모형을 결정하였다.

1. **탐색적 분석**

3-1) 변수선택 기준

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

[추정회귀계수]

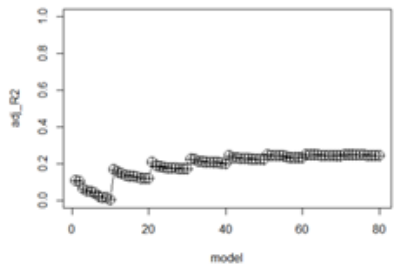
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.599883   0.053901  11.129  < 2e-16 ***
X1          -0.015983   0.003116  -5.128 3.84e-07 ***
X2          -0.004071   0.001734  -2.348  0.01916 *
X3           0.029770   0.006596   4.513 7.54e-06 ***
X4          -0.001433   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
```

모든 변수를 포함한 중회귀모형의 추정회귀계수를 구해보았다. p_value값을 확인해보았을 때 L_OUT_FQ(X1), BS3_1(X2), BP1(X3), LF_S4 (X6), DI2_2(X8), BE3_31(X10)변수가 유의수준 5%보다 작으므로 유의하고 BP16_1(X4), BP16_2(X5), DI1_2 (X7), DJ4_3(X9)변수가 유의하지 않은 것으로 생각된다.

[수정된 결정계수 adj_R]

수정된 결정계수는 71번째 모형에서 가장 큰 값이 도출되었고 null model과 비교했을 때 p_value가 유의수준 0.05보다 작으므로 귀무가설을 기각했다.

이때의 최종모형은 $Yi=\beta 0+\beta 1Xi1+\beta 2Xi2+\beta 3Xi3+\beta 5Xi5+\beta 6Xi6+ \beta 8Xi8+ \beta 9Xi9+\beta 10Xi10 +\epsilon i$ 이다.

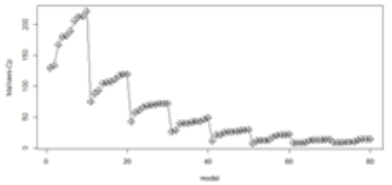



```
Model 1: EQ5D ~ L_OUT_FQ + BS3_1 + BP1 + LF_S4 + DI2_2 + BE3_31
Model 2: EQ5D ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    669 11.259
2    675 15.108 -6   -3.8485 38.111 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[Mallows-Cp]

Mallows-Cp는 51번째 모형에서 가장 작은 값이 도출되었고 null model과 비교했을 때 p_value가 유의수준 0.05보다 작으므로 귀무가설을 기각했다.

이때의 최종모형은 $Yi=\beta0+\beta1Xi1+\beta2Xi2+\beta3Xi3+\beta6Xi6+\ \beta8Xi8+\ \beta10\beta10Xi10\ +\epsilon i$이다.



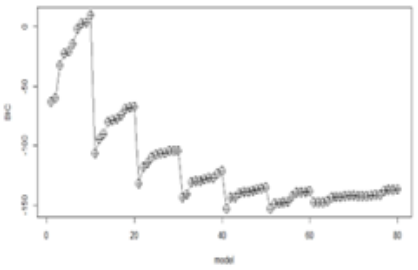
```
Model 1: EQ5D ~ L_OUT_FQ + BS3_1 + BP1 + BP16_2 + LF_S4 + DI2_2 + D34_3 +
    BE3_31
Model 2: EQ5D ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    667 11.219
2    675 15.108 -8   -3.8894 28.905 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[BIC]

BIC는 51번째 모형에서 가장 작은 값이 도출되었고 null model과 비교했을 때 p_value가 유의수준 0.05보다 작으므로 귀무가설을 기각했다.

이때의 최종모형은 $Yi=\beta0+\beta1Xi1+\beta2Xi2+\beta3Xi3+\beta6Xi6+\ \beta8Xi8+\ \beta10\beta10Xi10\ +\epsilon i$이다.

부분F검정, 수정된 결정계수, Mallows-Cp, BIC를 종합한 결과 X1, X2, X3, X6, X8, X10변수가 모형에 적합하다고 판단하였다.




```
Model 1: EQ5D ~ L_OUT_FQ + BS3_1 + BP1 + BP16_2 + LF_S4 + DI2_2 + D34_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
```

3-2) 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
   0.662185    -0.017859     0.057101     0.010837     0.029470    0.007410
     BS3_1
  -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
   0.662185    -0.017859     0.057101     0.010837     0.029470     0.007410
      BS3_1
  -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
   0.662185    -0.017859    -0.004323     0.029470     0.057101     0.007410
     BE3_31
   0.010837
```

세가지 경우 모두 X1, X2, X3, X6, X7, X10변수가 모형에 적합하다는 결과가 나왔다.

3-3) 회귀모형 진단 및 수정

[부분F검정]

1. 변수선택기준 결과를 바탕으로 Reduced_model1에서는 X4, X5, X7, X9 변수에 대해서 반응변수EQ-5D의 효과가 없다는 가설검정을 부분 F검정으로 살펴보았다.

가설 : $H0$ ： $\beta 4 = \beta 5 = \beta 7 = \beta 9 = 0$ 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
```

검정결과 p_value가 0.4652로써 유의수준 0.05에서 귀무가설을 기각하지 못했다. 따라서 BP16_1, BP16_2, DI1_2 , DJ4_3를 제외한 설명변수 6개의 축소모형이 10개의 변수를 모두 포함한 Full model보다 더 적절하다고 볼 수 있다.

1. 변수선택방법 결과를 바탕으로 Reduced_model1에서는 X4, X5, X8, X9 변수에 대해서 반응변수EQ-5D의 효과가 없다는 가설검정을 부분 F검정으로 살펴보았다.

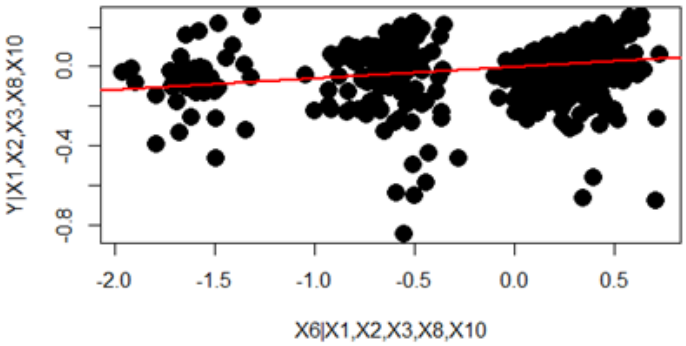
가설 : $H0$ ： $\beta 4 = \beta 5 = \beta 8 = \beta 9 = 0$ VS Ha ： $not\ H0$

```
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 **
```

검정결과 p_value가 0.007652로써 유의수준 0.05보다 작으므로 귀무가설을 기각했었다. X8 대신 X7변수를 Reduced model2에 넣으면 Full model이 적합하다는 결과가 나왔다. 따라서 Reduced model1이 더 적합하다고 가정했고 X1, X2, X3, X6, X8, X10을 적합한 변수라고 판단했다.

[편회귀그림]

X1, X2, X3, X6, X8, X10에 대한 편회귀그림을 그리고 편회귀 기울기를 구해보았다.



가장 뚜렷한 선형관계를 보인 변수는 X6이였다. 즉, 순수한 X6의 값이 반응변수 Y에 가장 많은 영향을 주고 있는 것을 확인하였다.

[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 귀무가설이 기각되므로 S^2이 과대추정되었다. 따라서 적합하지 않은 변수로 판단하였다.

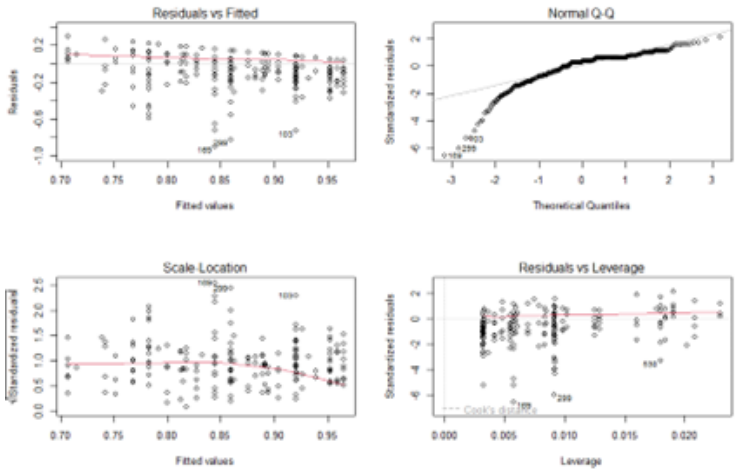
```
Model 1: Y ~ X2
Model 2: Y ~ factor(X2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    674 14.774
2    672 14.765  2 0.0086928 0.1978 0.8206
```

```
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
```

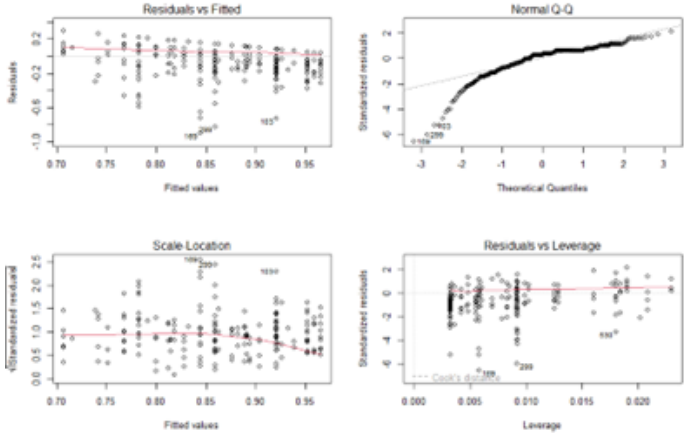
X2, X6, X8의 적합결여검정 결과 귀무가설이 기각되지않으므로 S^2이 과대추정되지 않았다. 따라서 적합한 변수로 판단하였다.

[변수변환]




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.672493   0.029802  22.565  < 2e-16
X2          -0.006375   0.001778  -3.585 0.000362
X6           0.076327   0.009650   7.910 1.06e-14
X8           0.008749   0.001973   4.435 1.08e-05
```

X2, X6, X8을 변수로 가진 Reduced3_model을 plot함수로 살펴본 결과 잔차는 중심에 모여있고 Q-Q plot은 완벽한 선형성을 보이지는 않았다. 따라서 변수변환이 필요하다고 판단한 후 √Y 변환을 해주었다.




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.816767   0.018031  45.299  < 2e-16
X2          -0.003706   0.001076  -3.444 0.000608
X6           0.041911   0.005844   7.172 1.97e-12
X8           0.005513   0.001194   4.619 4.62e-06
```

변수변환 결과 기존의 모형과 큰 차이가 없는 것으로 나타났다. 따라서 변수 변환 전 모형인 Reduced3_model을 최종모형 후보로 정했다.

1. **회귀 모형 선택**

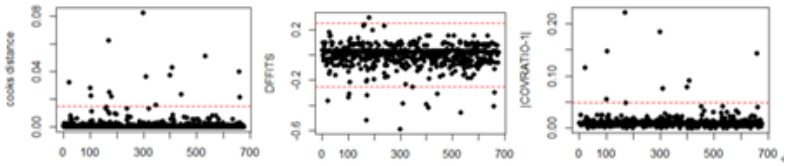
4-1) 모형 검정

[다중공선성 탐색]

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

다중공선성이란 설명변수들간의 상관관계가 높아 최소제곱추정량의 계산이 불가능해지는 것을 의미한다. 여기서는 VIF값이 5보다 작으므로 다중공선성이 없다.

[영향력 측도]




```
> influence.measures(Reduced3_model)
Influence measures of
         lm(formula = Y ~ X2 + X6 + X8)
```

영향력 측도들을 통해서 이상치를 구해본 결과 6, 22, 25, 29, 65, 82, 84, 100, 103, 104, 107, 109, 138, 157, 159, 162, 169, 170, 172, 174, 179, 192, 201, 238, 251, 252, 299,300, 308, 310, 312, 317, 320, 331, 335, 339, 347, 350, 363, 364, 380, 400, 406, 441, 452, 453, 463, 497, 525, 529, 530, 536, 539, 546, 556, 570, 581, 588, 634, 639, 650, 657, 660이 이상치로 도출되었다.

676개의 행으로 구성되었던 Reduced3_model에서 63개의 이상치를 제외하고 613개의 행으로

구성된 final_model을 최종모형으로 결정하였다.

[더빈-왓슨 검정]

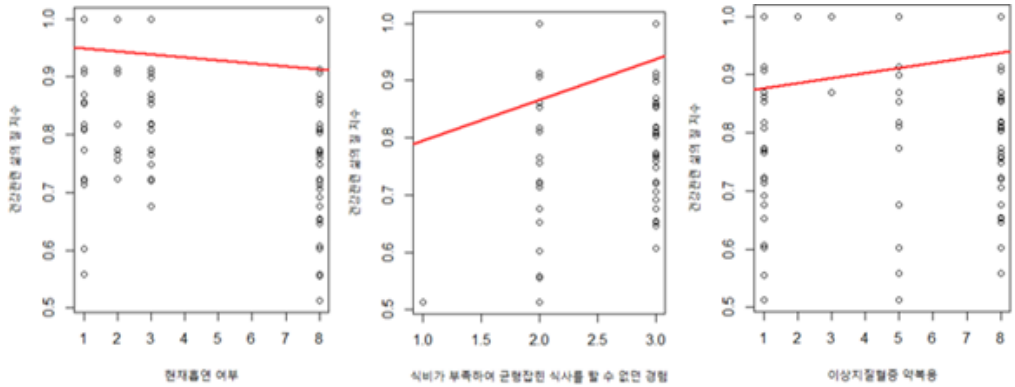
```
        Durbin-Watson test

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

더빈-왓슨 검정통계량이 0~4사이에 있으므로 오차항의 독립을 만족한다.

4-2) 최종 모형 결정

훈련자료(70%, 429개)와 확인자료(30%, 184개)로 구분한 후 예측오차를 구하고, PRESS와 R^2pedict를 구해보았다. PRESS와 SSE의 값이 비슷하고 R^2와 R^2predict가 비슷하므로 최종모형은 타당하다고 볼 수 있다.

```
> #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
```



최종모형의 설명변수 X2, X6, X8과 반응변수Y의 회귀모형을 그려봤다. 흡연 횟수가 늘어날수록 건강관련 삶의 질 지수가 낮아졌으며 식비가 부족하여 균형잡힌 식사를 할 수 없던 경험이 줄어들수록 건강관련 삶의 질 지수가 높아졌다. 또한 이상지질혈증 약복용 횟수가 줄어들수록 건강관련 삶의 질 지수가 높아지는 것을 확인할 수 있었다.

▼ III 결론

본 프로젝트를 통해서 1인 가구들의 건강 관련 삶의 질에 영향을 주고 있는 요인들을 파악해보았다. 탐색적 분석 후 최종 회귀모형을 검정한 결과 현재 흡연 여부, 식비가 부족하여 균형잡힌 식사를 할 수 없던 경험, 이상 지지혈증 약 복용이 1인 가구의 건강 관련 삶에 영향을 준다는 것을 확인하였다. 이러한 결과를 바탕으로 1인 가구의 건강에 대한 자료를 제공하였고 1인 가구의 건강 관련 삶의 질을 향상하기 위한 연구가 필요하다고 전하고 싶다.

이번 프로젝트를 진행하면서 아쉬웠던 점은 두 가지 이다. 첫째, 선형성, 독립성, 다중공선성 등을 확인하고 회귀진단을 실시하면서 유의미한 변수들이 감소하게 되었고, 이를 보면서 예비모형을 구축할 때 더욱 많은 설명변수를 관측해야 했음을 느꼈다. 둘째, 최종 회귀모형을 확인해보면 X2 변수와 X8 변수에서 8 비해당 부분에 자료가 많이 몰려있었다. 처음 국민건강영양조사 자료를 봤을 때 비해당 부분 또한 다른 답변과 비슷하게 구성될 것이라 생각해서 하나의 값으로 여기고 결측치 처리를 하지 않았었는데 이 또한 모름, 무응답과 함께 결측치 처리를 해야 했음을 느꼈다. 다음 프로젝트를 진행할 때는 이러한 한계점을 보완하며 더 정확하고 유의미한 회귀모형을 만들 것이다.

▼ IV 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)”

▼ 코드(R)

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/ca8060b7-8095-49cb-99af-8d07185d474c/%ED%9A%8C%EA%B7%80%EB%B6%84%EC%84%9D%EA%B8%B0%EB%A7%90%ED%94%84%EB%A1%9C%EC%A0%9D%ED%8A%B8%EA%B2%BD%EC%A0%9C%ED%95%99%EA%B3%BC20200856%EC%A0%95%EA%B0%80%EC%97%B0%EC%BD%94%EB%93%9C.r

```
library(styler)
library(reprex)
library(dplyr)
library(leaps)
library(MASS)
library(car)
library(lmtest)
library(qpcR)
setwd('C:/Users/SAMSUNG/Desktop/FIND-A/자유프로젝트')
dat <- read.csv('Hn19_all.csv',header=T)
model <- dat[,c('EQ5D','cfam','L_OUT_FQ','BS3_1','BP1','BP16_1','BP16_2',
               'LF_S4','DI1_2','DI2_2','DJ4_3','BE3_31')]
#데이터 처리전 표본 개수
nrow(model)
#모름,무응답은 결측치로 간주하고 제거
model = model %>% filter(cfam != 9)
model = model %>% filter(L_OUT_FQ != 9)
model = model %>% filter(BS3_1 != 9)
model = model %>% filter(BP1 != 8)
model = model %>% filter(BP1 != 9)
model = model %>% filter(BP16_1 != 88)
model = model %>% filter(BP16_1 != 99)
model = model %>% filter(BP16_2 != 88)
model = model %>% filter(BP16_2 != 99)
model = model %>% filter(LF_S4 != 4)
model = model %>% filter(LF_S4 != 9)
model = model %>% filter(DI1_2 != 9)
model = model %>% filter(DI2_2 != 9)
model = model %>% filter(DJ4_3 != 9)
model = model %>% filter(BE3_31 != 88)
model = model %>% filter(BE3_31 != 99)
#Missing Value(결측치 제거)
model <- na.omit(model)
# 전처리 후 데이터 개수
nrow(model)
```

```
#3.탐색적 분석

#3-1) 변수선택 기준
model = model %>% filter(cfam == 1)
nrow(model)

#변수 정의
Y <- model$EQ5D
X1 <- model$L_OUT_FQ
X2 <- model$BS3_1
X3 <- model$BP1
X4 <- model$BP16_1
X5 <- model$BP16_2
X6 <- model$LF_S4
X7 <- model$DI1_2
```

```
X8 <- model$DI2_2
X9 <- model$DJ4_3
X10 <- model$BE3_31

#Full_model
Full_model <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+BP16_1+BP16_2+LF_S4+DI1_2+DI2_2+DJ4_3+BE3_31, data = m
odel)
Full_model <- lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10)
summary(Full_model)

#[adj_R2]
model_reg <- model[,c('EQ5D','L_OUT_FQ','BS3_1','BP1','BP16_1','BP16_2',
                      'LF_S4','DI1_2','DI2_2','DJ4_3','BE3_31')]
regfit_sel <- regsubsets(x=EQ5D~.,data=model_reg,method = 'exhaustive',nbest=10)
summary(regfit_sel)
result_regfit <- summary(regfit_sel)
result_regfit$adjr2

#plot of adj_R2
plot(result_regfit$adjr2,ylim=c(0,1),pch=10,cex=2,ylab="adj_R2",xlab="model",type="b")

# final model
final_model_R2 <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+LF_S4+DI2_2+BE3_31, data = model_reg)
summary(final_model_R2)

# comparison with null model using partial-F test
null_model_R2 <- lm(EQ5D ~ 1,data=model_reg)
anova(final_model_R2,null_model_R2)

#[Mallows-cp]
result_regfit$cp

# plot of Mallows-cp
plot(result_regfit$cp,pch=10,cex=2,ylab="Mallows-Cp",xlab="model",type="b")

# final model
final_model_cp <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+BP16_2+LF_S4+DI2_2+DJ4_3+BE3_31, data = model_reg)
summary(final_model_cp)

# comparison with null model using partial-F test
null_model_cp <- lm(EQ5D ~ 1,data=model_reg)
anova(final_model_cp,null_model_R2)

#[BIC]
result_regfit$bic

# plot of BIC
plot(result_regfit$bic,pch=10,cex=2,ylab="BIC",xlab="model",type="b")

# final model
final_model_bic <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+BP16_2+LF_S4+DI2_2+DJ4_3+BE3_31, data = model_re
g)
summary(final_model_bic)

# comparison with null model using partial-F test
null_model_bic <- lm(EQ5D ~ 1,data=model_reg)
anova(final_model_bic,null_model_R2)
```

```
#3.탐색적 분석

#3-1) 변수선택 기준
model = model %>% filter(cfam == 1)
nrow(model)
#변수 정의
```

```
Y <- model$EQ5D
X1 <- model$L_OUT_FQ
X2 <- model$BS3_1
X3 <- model$BP1
X4 <- model$BP16_1
X5 <- model$BP16_2
X6 <- model$LF_S4
X7 <- model$DI1_2
X8 <- model$DI2_2
X9 <- model$DJ4_3
X10 <- model$BE3_31
#Full_model
Full_model <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+BP16_1+BP16_2+LF_S4+DI1_2+DI2_2+DJ4_3+BE3_31, data = mode
Full_model <- lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10)
summary(Full_model)
#[adj_R2]
model_reg <- model[,c('EQ5D','L_OUT_FQ','BS3_1','BP1','BP16_1','BP16_2',
                      'LF_S4','DI1_2','DI2_2','DJ4_3','BE3_31')]
regfit_sel <- regsubsets(x=EQ5D~.,data=model_reg,method = 'exhaustive',nbest=10)
summary(regfit_sel)
result_regfit <- summary(regfit_sel)
result_regfit$adjr2
#plot of adj_R2
plot(result_regfit$adjr2,ylim=c(0,1),pch=10,cex=2,ylab="adj_R2",xlab="model",type="b")
# final model
final_model_R2 <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+LF_S4+DI2_2+BE3_31, data = model_reg)
summary(final_model_R2)
# comparison with null model using partial-F test
null_model_R2 <- lm(EQ5D ~ 1,data=model_reg)
anova(final_model_R2,null_model_R2)
#[Mallows-cp]
result_regfit$cp
# plot of Mallows-cp
plot(result_regfit$cp,pch=10,cex=2,ylab="Mallows-Cp",xlab="model",type="b")
# final model
final_model_cp <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+BP16_2+LF_S4+DI2_2+DJ4_3+BE3_31, data = model_reg)
summary(final_model_cp)
# comparison with null model using partial-F test
null_model_cp <- lm(EQ5D ~ 1,data=model_reg)
anova(final_model_cp,null_model_R2)
#[BIC]
result_regfit$bic
# plot of BIC
plot(result_regfit$bic,pch=10,cex=2,ylab="BIC",xlab="model",type="b")
# final model
final_model_bic <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+BP16_2+LF_S4+DI2_2+DJ4_3+BE3_31, data = model_reg)
summary(final_model_bic)
# comparison with null model using partial-F test
null_model_bic <- lm(EQ5D ~ 1,data=model_reg)
anova(final_model_bic,null_model_R2)
```

```
#3-2) 변수선택 방법

#[stepwise]
null_model <- lm(EQ5D ~ 1,data=model)
step(null_model, scope = ~ L_OUT_FQ+BS3_1+BP1+BP16_1+BP16_2+LF_S4+DI1_2+DI2_2+DJ4_3+BE3_31,directi
#[forward]
null_model <- lm(EQ5D ~ 1,data=model)
step(null_model, scope = ~ L_OUT_FQ+BS3_1+BP1+BP16_1+BP16_2+LF_S4+DI1_2+DI2_2+DJ4_3+BE3_31,directi
#[backward]
Full_model <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+BP16_1+BP16_2+LF_S4+DI1_2+DI2_2+DJ4_3+BE3_31, data = mode
step(Full_model,direction="backward",test="F")
```

```
#3-3) 회귀모형 진단 및 수정

#[부분F검정]
#Full_model
Full_model <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+BP16_1+BP16_2+LF_S4+DI1_2+DI2_2+DJ4_3+BE3_31, data = mode
Full_model <- lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10)
Y_hat_F <- predict(Full_model, newdata = data.frame(X1=X1,X2=X2,X3=X3,X4=X4,X5=X5,X6=X6,X7=X7,X8=X
SST <- sum((Y-mean(Y))^2);SST
SSR_F <- sum((Y_hat_F-mean(Y))^2);SSR_F
SSE_F <- sum((Y-Y_hat_F)^2);SSE_F
MSR_F <- SSR_F/(11-1);MSR_F
MSE_F <- SSE_F/(676-11);MSE_F
F0_F <- MSR_F/MSE_F;F0_F
pf(F0_F,df1=11-1,df2=676-11,lower.tail=F)
#Reduced_model - X1,X2,X3,X6,X8,X10 유의
Reduced1_model <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+LF_S4+DI2_2+BE3_31, data = model)
Reduced1_model <- lm(Y~X1+X2+X3+X6+X8+X10)
Y_hat_R <- predict(Reduced1_model, newdata = data.frame(X1=X1,X2=X2,X3=X3,X6=X6,X8=X8,X10=X10))
SST <- sum((Y-mean(Y))^2);SST
SSR_R <- sum((Y_hat_R-mean(Y))^2);SSR_R
SSE_R <- sum((Y-Y_hat_R)^2);SSE_R
MSR_R <- SSR_R/(7-1);MSR_R
MSE_R <- SSE_R/(676-7);MSE_R
F0_R <- MSR_R/MSE_R;F0_R
pf(F0_R,df1=7-1,df2=263-7,lower.tail=F)
#Calculation of F0
R_b2b1 <- SSE_R - SSE_F ; R_b2b1
F0 <- (R_b2b1/(11-7))/(SSE_F/(676-11)) ; F0
pf(F0,df1=11-7,df2=676-11,lower.tail=F)
anova(Reduced1_model,Full_model)
#Reduced_model2 - X1,X2,X3,X6,X7,X10 유의
Reduced2_model <- lm(EQ5D~L_OUT_FQ+BS3_1+BP1+LF_S4+DI1_2+BE3_31, data = model)
Reduced2_model <- lm(Y~X1+X2+X3+X6+X7+X10)
summary(Reduced1_model)
Y_hat_R <- predict(Reduced1_model, newdata = data.frame(X1=X1,X2=X2,X3=X3,X6=X6,X7=X7,X10=X10))
SST <- sum((Y-mean(Y))^2);SST
SSR_R <- sum((Y_hat_R-mean(Y))^2);SSR_R
SSE_R <- sum((Y-Y_hat_R)^2);SSE_R
MSR_R <- SSR_R/(7-1);MSR_R
MSE_R <- SSE_R/(676-7);MSE_R
F0_R <- MSR_R/MSE_R;F0_R
pf(F0_R,df1=7-1,df2=263-7,lower.tail=F)
#Calculation of F0
R_b2b1 <- SSE_R - SSE_F ; R_b2b1
F0 <- (R_b2b1/(11-7))/(SSE_F/(676-11)) ; F0
pf(F0,df1=11-7,df2=676-11,lower.tail=F)
anova(Reduced2_model,Full_model)
#[편회귀그림]
#X1의 편회귀그림
model236810 <- lm(Y~X2+X3+X6+X8+X10);model236810
modelX236810 <- lm(X1~X2+X3+X6+X8+X10);modelX236810
y.X2X3X6X8X10 <- resid(model236810)
X1.X2X3X6X8X10 <- resid(modelX236810)
plot(X1.X2X3X6X8X10,y.X2X3X6X8X10,pch=16,cex=2,xlab="X1|X2,X3,X6,X8,X10",ylab="Y|X2,X3,X6,X8,X10")
model_parX1 <- lm(y.X2X3X6X8X10~X1.X2X3X6X8X10) ; model_parX1
abline(model_parX1,col='red',lwd=2)
cor.test(X1.X2X3X6X8X10,y.X2X3X6X8X10)
#X2의 편회귀그림
model136810 <- lm(Y~X1+X3+X6+X8+X10);model136810
modelX136810 <- lm(X2~X1+X3+X6+X8+X10);modelX136810
y.X1X3X6X8X10 <- resid(model136810)
X2.X1X3X6X8X10 <- resid(modelX136810)
plot(X2.X1X3X6X8X10,y.X1X3X6X8X10,pch=16,cex=2,xlab="X2|X1,X3,X6,X8,X10",ylab="Y|X1,X3,X6,X8,X10")
model_parX2 <- lm(y.X1X3X6X8X10~X2.X1X3X6X8X10) ; model_parX2
abline(model_parX2,col='red',lwd=2)
```

```
cor.test(X2.X1X3X6X8X10,y.X1X3X6X8X10)
#X3의 편회귀그림
model126810 <- lm(Y~X1+X2+X6+X8+X10);model126810
modelX126810 <- lm(X3~X1+X2+X6+X8+X10);modelX126810
y.X1X2X6X8X10 <- resid(model126810)
X3.X1X2X6X8X10 <- resid(modelX126810)
plot(X3.X1X2X6X8X10,y.X1X2X6X8X10,pch=16,cex=2,xlab="X3|X1,X2,X6,X8,X10",ylab="Y|X1,X2,X6,X8,X10")
model_parX3 <- lm(y.X1X2X6X8X10~X3.X1X2X6X8X10) ; model_parX3
abline(model_parX3,col='red',lwd=2)
cor.test(X3.X1X2X6X8X10,y.X1X2X6X8X10)
#X6의 편회귀그림
model123810 <- lm(Y~X1+X2+X3+X8+X10);model123810
modelX123810 <- lm(X6~X1+X2+X3+X8+X10);modelX123810
y.X1X2X3X8X10 <- resid(model123810)
X6.X1X2X3X8X10 <- resid(modelX123810)
plot(X6.X1X2X3X8X10,y.X1X2X3X8X10,pch=16,cex=2,xlab="X6|X1,X2,X3,X8,X10",ylab="Y|X1,X2,X3,X8,X10")
model_parX6 <- lm(y.X1X2X3X8X10~X6.X1X2X3X8X10) ; model_parX6
abline(model_parX6,col='red',lwd=2)
cor.test(X6.X1X2X3X8X10,y.X1X2X3X8X10)
#X8의 편회귀그림
model123610 <- lm(Y~X1+X2+X3+X6+X10);model123610
modelX123610 <- lm(X8~X1+X2+X3+X6+X10);modelX123610
y.X1X2X3X6X10 <- resid(model123610)
X8.X1X2X3X6X10 <- resid(modelX123610)
plot(X8.X1X2X3X6X10,y.X1X2X3X6X10,pch=16,cex=2,xlab="X8|X1,X2,X3,X6,X10",ylab="Y|X1,X2,X3,X6,X10")
model_parX8 <- lm(y.X1X2X3X6X10~X8.X1X2X3X6X10) ; model_parX8
abline(model_parX8,col='red',lwd=2)
cor.test(X8.X1X2X3X6X10,y.X1X2X3X6X10)
#X10의 편회귀그림
model12368 <- lm(Y~X1+X2+X3+X6+X8);model12368
modelX12368 <- lm(X10~X1+X2+X3+X6+X8);modelX12368
y.X1X2X3X6X8 <- resid(model12368)
X10.X1X2X3X6X8 <- resid(modelX12368)
plot(X10.X1X2X3X6X8,y.X1X2X3X6X8,pch=16,cex=2,xlab="X10|X1,X2,X3,X6,X8",ylab="Y|X1,X2,X3,X6,X8")
model_parX10 <- lm(y.X1X2X3X6X8~X10.X1X2X3X6X8) ; model_parX10
abline(model_parX10,col='red',lwd=2)
cor.test(X10.X1X2X3X6X8,y.X1X2X3X6X8)
#[적합결여검정]
# H0 : E(y) = b0+b1*x1 VS Ha : E(y) != b0+b1*x1 ->적합X
fit.lm1 <- lm(Y ~ X1)
fit.pe1 <- lm(Y ~ factor(X1))
anova(fit.lm1,fit.pe1)
# H0 : E(y) = b0+b2*x2 VS Ha : E(y) != b0+b2*x2 -> 적합O
fit.lm2 <- lm(Y ~ X2)
fit.pe2 <- lm(Y ~ factor(X2))
anova(fit.lm2,fit.pe2)
# H0 : E(y) = b0+b3*x3 VS Ha : E(y) != b0+b3*x3 -> 적합X
fit.lm3 <- lm(Y ~ X3)
fit.pe3 <- lm(Y ~ factor(X3))
anova(fit.lm3,fit.pe3)
# H0 : E(y) = b0+b6*x6 VS Ha : E(y) != b0+b6*x6 -> 적합O
fit.lm6 <- lm(Y ~ X6)
fit.pe6 <- lm(Y ~ factor(X6))
anova(fit.lm6,fit.pe6)
# H0 : E(y) = b0+b8*x8 VS Ha : E(y) != b0+b8*x8 -> 적합O
fit.lm8 <- lm(Y ~ X8)
fit.pe8 <- lm(Y ~ factor(X8))
anova(fit.lm8,fit.pe8)
# H0 : E(y) = b0+b10*x10 VS Ha : E(y) != b0+b10*x10 -> 적합X
fit.lm10 <- lm(Y ~ X10)
fit.pe10 <- lm(Y ~ factor(X10))
anova(fit.lm10,fit.pe10)
#Reduced3_model
Reduced3_model <- lm(Y ~ X2 + X6 + X8)
summary(Reduced3_model)
par(mfrow=c(2,2))
```

```
plot(Reduced3_model)
#[변수변환]
# √Y transformation
Y_sqrt <- sqrt(Y)
Reduced3_model_sqrt <- lm(Y_sqrt ~ X2 + X6 + X8)
summary(Reduced3_model_sqrt)
par(mfrow=c(2,2))
plot(Reduced3_model_sqrt)
```

```
#4.회귀 모형 선택

#4-1) 모형 검정

#[다중공선성 탐색]
vif(Reduced3_model)
#[영향력 측도]
#Cook's distance
par(mfrow=c(1,1))
cooks.distance(Reduced3_model)
tail(Reduced3_model)
n <- 250
p <- 4
cook_standard <- 3.67 / (n-p) ; cook_standard
plot(cooks.distance(Reduced3_model),pch=16,ylab="cooks distance")
abline(h=cook_standard,col="red",lty=2)
#DFFITS
DFFITS_standard <- 2*sqrt(p/n);DFFITS_standard
plot(dffits(Reduced3_model),pch=19,ylab="DFFITS")
abline(h=DFFITS_standard,col="red",lty=2)
abline(h=-DFFITS_standard,col="red",lty=2)
#COVRATIOS
COVRATIO_standard <- 3*p/n;COVRATIO_standard
plot(abs(covratio(Reduced3_model)-1),pch=19,ylab="|COVRATIO-1|")
abline(h=COVRATIO_standard,col="red",lty=2)
#전체 영향력 측도 확인
options(max.print=1000000)
influence.measures(Reduced3_model)
#이상치 제외한 model
final_model <- model[-c(6, 22, 25, 29, 65, 82, 84, 100, 103, 104, 107, 109, 138, 157, 159, 162, 16
                        172, 174, 179, 192, 201, 238, 251, 252, 299,300, 308, 310, 312, 317, 32
                        335, 339, 347, 350, 363, 364, 380, 400, 406, 441, 452, 453, 463, 497, 5
                        536, 539, 546, 556, 570, 581, 588, 634, 639, 650, 657, 660),]
Final_model <- lm(Y ~ X2 + X6 + X8)
#[더빈-왓슨 검정]
dwtest(Final_model)
```

```
#4-2)최종 모형 결정
set.seed(1234)
rn <- sample(x=c(1:613), size = 613, replace=F);rn
final_model$rn <- rn
train_dat <- final_model[final_model$rn>184,]
test_dat <- final_model[final_model$rn<=184,]
dim(train_dat)
dim(test_dat)
#predicted error
train_model <- lm(EQ5D~BS3_1+LF_S4+DI2_2,data=train_dat);summary(train_model)
predict_value <- predict(train_model,newdata = test_dat[,c('BS3_1','LF_S4','DI2_2')])
predict_error <- sum((test_dat$sales-predict_value)^2);predict_error
#PRESS
last_model <- lm(EQ5D~BS3_1+LF_S4+DI2_2,data=final_model)
PRESS_last <- PRESS(last_model)
PRESS_last$stat
```

```
#SST, SSE, SSR
SST <- sum((final_model$EQ5D-mean(final_model$EQ5D))^2)
SSE <- sum(resid(last_model)^2)
SSR <- SST-SSE
#PRESS VS SSE
PRESS_last$stat
SSE
#R2 VS R2_predic
1-(SSE/SST)
1-(PRESS_last$stat/SST)

#최종모형
plot(x=final_model$BS3_1,y=final_model$EQ5D,xlab='현재흡연 여부',ylab='건강관련 삶의 질 지수')
model2 <- lm(final_model$EQ5D~final_model$BS3_1)
abline(model2, col='red', lwd=2)

plot(x=final_model$LF_S4,y=final_model$EQ5D,xlab='식비가 부족하여 균형잡힌 식사를 할 수 없던 경험',ylab='건강
model2 <- lm(final_model$EQ5D~final_model$LF_S4)
abline(model2, col='red', lwd=2)

plot(x=final_model$DI2_2,y=final_model$EQ5D,xlab='이상지질혈증 약복용',ylab='건강관련 삶의 질 지수')
model2 <- lm(final_model$EQ5D~final_model$DI2_2)
abline(model2, col='red', lwd=2)
```