2022-2 회귀분석(1139301-01)

Final project

학과: AI빅데이터융합경영학과

학번: 20212548

이름: 김지은

I 서론

- 국민 건강영양조사는 국민의 건강행태, 만성질환 유병현황, 식품 및 영양섭취실태에 관한 조사로, 국가 단위의 대표성과 신뢰성을 갖춘 통계를 산출한다. 이를 통해 국민건강증진종합계획의 목표 설정 및 평가, 건강증진 프로그램 개발 등 보건정책의 기초자료로 활용하는 것이다.

- 기존에 제출했던 연구 계획서에는 스트레스와 음주, 흡연의 상관관계였다. 하지만 종속변수를 범주형 변수로 설정하고 진행한 결과, 분석 결과가 잘 나오지 않아, '행복'이라는 주제로 바꾸게 되었다. 하지만 행복 또한 범주형 변수이다 보니, 분석하는데 어려움이 생길 것이라는 판단을 하게 되었고, 논문을 통해 행복과 비례한 변수를 찾아 종속변수로 대체하였다. 종속변수와 독립변수가 양의 상관관계를 가지면 행복도 그에 해당하는 독립변수와 영향이 있다는 것으로 판단한다. 최적의 모형을 통해 우리나라 사람들의 행복결정요인에 대해 알아보고, 행복향상방안에 대해 알아본다.[1]

II 본론

- 행복결정요인에 영향을 미칠 것 같은 변수들을 선정하여 1차 모형을 설정 후, 변수 선택 및 변수 변환 과정을 반복해 최종 모형을 선정한다.

<표 3-110> 행복지수 산출을 위한 지표

영역		우선순위	영향성도
심리적 안정	① 자신에 대한 자아 존중감 정도	2.4	8.0
	② 자신에 대한 긍정적인 가치관 및 감정	2.8	7.9
	③ 현재 자신의 모습에 대한 만족도	2.9	7.9
가족·결혼	④ 현재의 가족(결혼)생활에 대한 만족도	1.5	8.7
	⑤ 현재의 부부생활(이성교제)에 대한 만족도	2.1	8.1
	⑥ 출산 및 자녀성장에 대한 만족도	2.8	7.4
개인적 관계	⑦ 가족원 관계에 대한 만족도	1.6	8.5
	⑧ 친구 및 동료와 관계에 대한 만족도	2.2	7.9
	⑨ 타인에게 긍정적으로 인정받는 정도	2.4	7.6
지역사회	⑩ 지역사회환경에 대한 만족도	1.2	6.3
일상생활	⑪ 만족스러운 수면(양, 질)	2.1	7.6
	⑫ 여가 및 휴식에 대한 만족도	1.8	8.0
경제적 안정	⑬ 원하는 만큼 재화(소득)을 소유(관리)하고 있다는 충족감	1.9	7.4
	⑭ 원하는 것을 언제든 사거나 가질 수 있다는 유능감	1.9	7.4
일	⑮ 현재의 일 종류와 자신이 원하는 것과의 일치 정도	1.8	8.5
	⑯ 현재 일에 대한 급여 및 근무환경에 대한 만족도	2.5	8.1
	⑰ 자신의 일에 대한 보람 정도	2.4	8.1
건강	⑱ 자신의 주관적인 건강수준	1.6	8.4
	⑲ 가족의 건강수준	1.8	8.7

- 논문을 찾아본 결과, 건강수준, 소득 수준 등이 대표적으로 행복지수에 비례한다는 결론을 통해 수치형 변수 중, 수면과 월평균 가구총소득을 종속변수로 설정하였다.

- 설명변수의 선정방식의 경우, 크게 식사, 음주, 흡연, 건강의 카테고리로 나누어 수치형 변수를 가지는 변수들로 변수를 설정하였다. 최근 1년 동안 1주 동안 아침식사 빈도(L_BR_FQ), 한 번에 마시는 음주량_잔(BD2_14), 평소 하루 앉아서 보내는 시간(BE8_1), 하루평균 흡연량(BS3_2), 1일 당 섭취량(N_SUGAR), 1일 탄수화물 섭취량(N_CHO), 걷기 지속 시간(BE3_32)을 예비 모델의 변수로 채택하였

[1] https://repository.kihasa.re.kr/bitstream/201002/544/1/%EC%97%B0%EA%B5%AC%EB%B3%B4%EA%B3%A0%EC%84%9C%202008-13.PDF

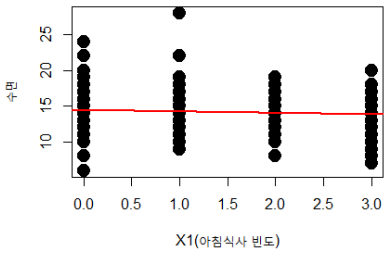
다.

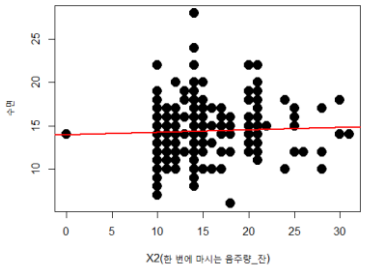
- 종속변수 - 1) 수면, 2) 월평균 가구총소득(ainc)

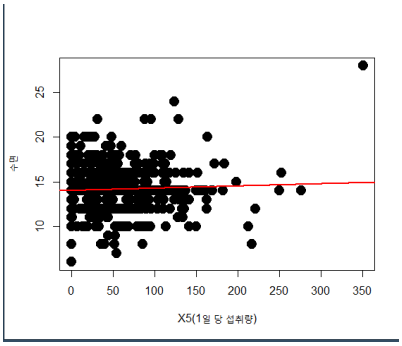
- 독립변수- 최근 1년 동안 1주 동안 아침식사 빈도(L_BR_FQ), 한 번에 마시는 음주량_잔(BD2_14), 평소 하루 앉아서 보내는 시간(BE8_1), 하루평균 흡연량(BS3_2), 1일 당 섭취량(N_SUGAR), 1일 탄수화물 섭취량(N_CHO), 걷기 지속 시간(BE3_32), 총 7개의 변수를 택했으며, 분석대상자는 해당 변수들의 결측치를 제외한 788명, 786명이다.

종속변수1: 수면

- 여러 변수들 중, 한 번에 마시는 음주량_잔 및 1일 당 섭취량이 다른 독립변수들보다 양의 상관관계를 가진다는 것을 catter plot을 통해 확인할 수 있다.







X: 아침식사 빈도 X: 한 번에 마시는 음주량_잔 X: 1일 당 섭취량

#Full model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.1685974  0.1913352  74.051  < 2e-16 ***
X1          -0.2032719  0.0624049  -3.257  0.00117 **
X2           0.0596934  0.0161310   3.701  0.00023 ***
X3          -0.0472366  0.0215758  -2.189  0.02887 *
X4          -0.0268493  0.0121340  -2.213  0.02721 *
X5           0.0025997  0.0024020   1.082  0.27946
X6           0.0004250  0.0007629   0.557  0.57758
X7          -0.1437377  0.0796377  -1.805  0.07148 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.046 on 780 degrees of freedom
Multiple R-squared:  0.04088,   Adjusted R-squared:  0.03227
F-statistic: 4.749 on 7 and 780 DF,  p-value: 3.064e-05
```

Full model을 돌려본 결과, 1일 당 섭취량(N_SUGAR), 1일 탄수화물 섭취량(N_CHO), 걷기 지속 시간(BE3_32)의 유의성이 낮다고 판단된다.

#다중공선성

다중공선성의 경우, 큰 값이 특출나게 보이지는 않았다.

```
> vif(Full_model)
      X1       X2       X3       X4       X5       X6       X7
1.300803 2.853631 2.389402 1.265554 2.026129 2.308203 1.130667
```

변수 선택

1) Adj-R2

```
 [1]  0.0133230024  0.0096505769  0.0013994943 -0.0002124993 -0.0006155824
 [6] -0.0007674861 -0.0011572168  0.0179057305  0.0175515419  0.0170627458
[11]  0.0134962081  0.0133733214  0.0133205010  0.0125939958  0.0218140845
[16]  0.0212200272  0.0211602572  0.0210220647  0.0194933234  0.0181568171
[21]  0.0180264507  0.0261175409  0.0255565393  0.0255507894  0.0252918324
[26]  0.0249828569  0.0245796767  0.0237003987  0.0300214328  0.0296049608
[31]  0.0291601979  0.0283773881  0.0280421493  0.0276380723  0.0271924177
[36]  0.0331259784  0.0320591910  0.0294740951  0.0275714100  0.0274438618
[41]  0.0203638079  0.0165425119  0.0322715236
```

앞서 진행한 것에 대해 간단한 최적 모형을 찾기 위해 변수 선택을 진행한다. 변수 선택 방법은 총 2가지(adj-R2, Mallow-cp)를 사용하였다.

분석 결과, 가장 높게 나온 값인 0.3312로 36번째의 모델을 택한다. 독립변수가 6개인 모델로 변수를 선택하였다.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.210747   0.175666  80.896  < 2e-16 ***
X1          -0.191894   0.058943  -3.256 0.001180 **
X2           0.060260   0.016092   3.745 0.000194 ***
X3          -0.048493   0.021448  -2.261 0.024039 *
X4          -0.026676   0.012125  -2.200 0.028089 *
X5           0.003532   0.001722   2.051 0.040579 *
X7          -0.148348   0.079172  -1.874 0.061338 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.045 on 781 degrees of freedom
Multiple R-squared:  0.0405,    Adjusted R-squared:  0.03313
F-statistic: 5.494 on 6 and 781 DF,  p-value: 1.386e-05
```

검정 결과, 5%의 경우, X7변수가 유의성을 띄지는 않지만, 1%를 기준으로 하면, 유의수준을 벗어나는 값이 존재하지는 않았다.

#null model과 adj-R2와의 비교

```
Model 1: Y ~ X1 + X2 + X3 + X4 + X5 + X7
Model 2: Y ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    781 3266.1
2    787 3403.9 -6   -137.85 5.4939 1.386e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

비교한 결과, adj-R2의 모델이 아주 유의한 것으로 나타났다.

2) Mallow-cp

```
 [1] 17.390203 20.372988 27.074611 28.383890 28.711279 28.834657 29.151201 14.653215 14.940525 15.337025 18.230122
[12] 18.329805 18.372652 18.961977 12.472038 12.953310 13.001733 13.113688 14.352190 15.434953 15.540569  9.979257
[23] 10.433169 10.437822 10.647347 10.897342 11.223560 11.934994  7.818249  8.154790  8.514194  9.146765  9.417664
[34]  9.744190 10.104313  6.310417  7.171362  9.257649 10.793201 10.896138 16.610057 19.694013  8.000000
```

Mallow-cp 분석 결과, 7번째 값이 29.1512로 가장 높게 나왔다. 따라서 평균 흡연량 (X4)을 하나의 변수로 갖는 모델을 선택한다.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.207135   0.083735 169.668   <2e-16 ***
X4          -0.003297   0.010971  -0.301    0.764
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.081 on 786 degrees of freedom
Multiple R-squared:  0.0001149, Adjusted R-squared:  -0.001157
F-statistic: 0.09032 on 1 and 786 DF,  p-value: 0.7638
```

하지만, 검정결과 0.764로 유의수준에 벗어난다. 따라서 이 모델은 사용하지 않는다. adj-R2을 통해 변수 X6을 제외한 나머지 변수 6개로 만들어진 모델을 선택한다.

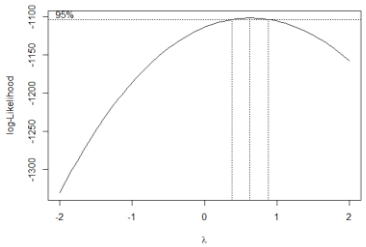
#단계별 회귀

```
Coefficients:
(Intercept)          X1          X5          X2          X3          X4          X7
  14.210747   -0.191894    0.003532    0.060260   -0.048493   -0.026676   -0.148348
```

단계별 회귀를 실시하여 adj-R2에서 선택된 변수 6개 중 더 유의미한 설명변수가 있는지 확인하기 위해 단계별회귀를 실시했다. 실시한 결과, 변수 6개 모두 유의미해 모두 선택하였다.

#변수 변환

#boxcox 변환



Boxcox 결과, lambda= 1에 가까운 값이 나왔다. 하지만 3가지 변수 변환을 다 실행해본다.

Y

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.210747   0.175666  80.896  < 2e-16 ***
X1          -0.191894   0.058943  -3.256 0.001180 **
X2           0.060260   0.016092   3.745 0.000194 ***
X3          -0.048493   0.021448  -2.261 0.024039 *
X4          -0.026676   0.012125  -2.200 0.028089 *
X5           0.003532   0.001722   2.051 0.040579 *
X7          -0.148348   0.079172  -1.874 0.061338 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.045 on 781 degrees of freedom
Multiple R-squared:  0.0405,    Adjusted R-squared:  0.03313
F-statistic: 5.494 on 6 and 781 DF,  p-value: 1.386e-05
```

log

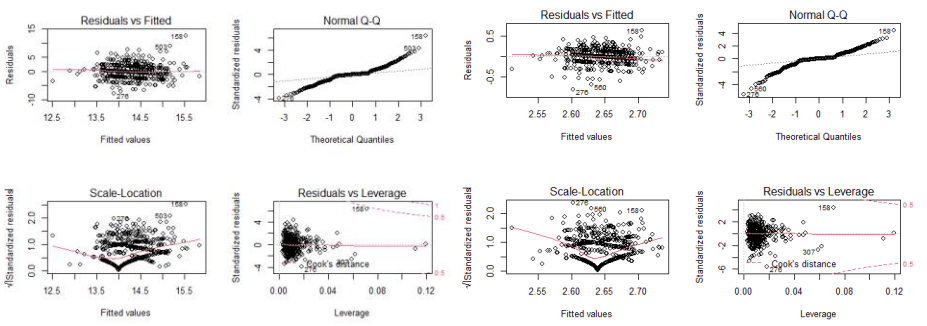
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.6563564  0.0126378 210.192  < 2e-16 ***
X1          -0.0133275  0.0042405  -3.143  0.00174 **
X2           0.0039352  0.0011577   3.399  0.00071 ***
X3          -0.0045339  0.0015430  -2.938  0.00340 **
X4          -0.0021728  0.0008723  -2.491  0.01295 *
X5           0.0001776  0.0001239   1.433  0.15217
X7          -0.0109004  0.0056958  -1.914  0.05601 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1471 on 781 degrees of freedom
Multiple R-squared:  0.03515,    Adjusted R-squared:  0.02774
F-statistic: 4.743 on 6 and 781 DF,  p-value: 9.309e-05
```

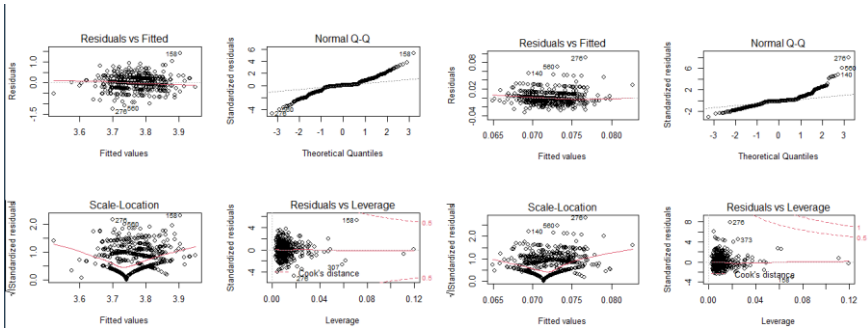
Sqrt

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.7720329  0.0233324 161.665  < 2e-16 ***
X1          -0.0251206  0.0078290  -3.209 0.001388 **
X2           0.0076557  0.0021373   3.582 0.000362 ***
X3          -0.0074054  0.0028488  -2.599 0.009513 **
X4          -0.0037759  0.0016104  -2.345 0.019294 *
X5           0.0003932  0.0002287   1.719 0.085966 .
X7          -0.0200308  0.0105158  -1.905 0.057168 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2716 on 781 degrees of freedom
Multiple R-squared:  0.03697,    Adjusted R-squared:  0.02957
F-statistic: 4.997 on 6 and 781 DF,  p-value: 4.903e-05
```

역수

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.005e-02  9.860e-04  71.043  < 2e-16 ***
X1           9.738e-04  3.308e-04   2.944  0.00334 **
X2          -2.725e-04  9.032e-05  -3.017  0.00263 **
X3           4.314e-04  1.204e-04   3.583  0.00036 ***
X4           1.887e-04  6.805e-05   2.773  0.00569 **
X5          -9.582e-06  9.666e-06  -0.991  0.32183
X7           8.179e-04  4.444e-04   1.841  0.06606 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01148 on 781 degrees of freedom
Multiple R-squared:  0.03662,    Adjusted R-squared:  0.02922
F-statistic: 4.948 on 6 and 781 DF,  p-value: 5.551e-05
```





R-squared의 경우, Y가 가장 높다. 다른 변환된 그림들 보다는 Y의 Residuals vs Fitted의 선이 가장 일직선에 가깝다고 판단할 수 있다.

#Final model 회귀 모형의 선택

```
> sum(resid(Final_model)^2)
[1] 16.90395
> sum((resid(Final_model)/(1-hatvalues(Final_model)))^2)
[1] 17.35178
```

```
> # R2 vs R2_predic
> 1-(SSE/SST) # R2 = 0.995034
[1] 0.995034
> 1-(press/SST) # R2_predic = 0.9949024
[1] 0.9949024
```

SSE VS PRESS

16.90395 VS 17.35178으로 차이가 크지 않아 예측의 정확도가 높다고 할 수 있다.

R2 vs R2-predic

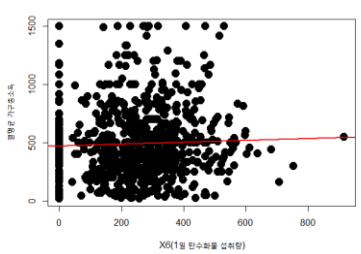
0.9950 VS 0.9949 값으로 차이가 거의 없어 예측의 정확도가 높다고 할 수 있고, 최종모형을 새로운 자료에 적용할 수 있다고 판단된다.

최종모형:

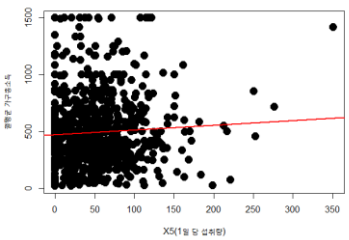
Y= -0.0133275 X1 + 0.0039352 X2 -0.0045339 X3 -0.0021728 X4 +0.0001776 X5 -0.0109004 X7

종속변수2: 월평균 가구총소득

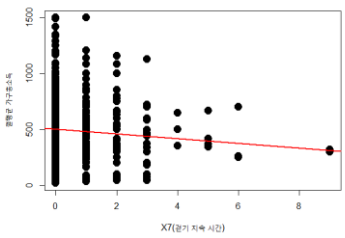
- Scatter plot을 그려본 결과, 1일 당 섭취량, 1일 탄수화물 섭취량, 걷기 지속시간이 다른 변수들보다 뚜렷한 상관관계 그래프가 나타난다.



X: 1일 당 섭취량



X: 1일 탄수화물 섭취량



X: 걷기 지속 시간

#Full model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 476.140876  29.775197  15.991  < 2e-16 ***
X1           -3.878841   9.707853  -0.400  0.68959
X2           -0.685123   2.505347  -0.273  0.78457
X3            6.608149   3.349963   1.973  0.04889 *
X4           -5.162009   1.883841  -2.740  0.00628 **
X5            0.455341   0.373130   1.220  0.22271
X6            0.001258   0.118474   0.011  0.99153
X7          -23.529339  12.363986  -1.903  0.05740 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 317.6 on 778 degrees of freedom
Multiple R-squared:  0.02257,   Adjusted R-squared:  0.01377
F-statistic: 2.566 on 7 and 778 DF,  p-value: 0.01282

>
```

Full model을 돌려본 결과, 평소 하루에 앉아서 보내는 시간 및 하루평균 흡연량의 제외한 나머지 변수들은 유의성이 낮다고 판단되었다.

#다중공선성

다중공선성의 경우, 큰 값이 특출나게 보이지는 않았다.

```
> vif(Full_model)
       X1        X2        X3        X4        X5        X6        X7
1.301832 2.846466 2.383512 1.264706 2.026215 2.308987 1.130119
```

변수 선택

- 앞서 진행한 것에 대해 간단한 최적 모형을 찾기 위해 변수 선택을 진행한다. 변수 선택 방법은 총 2가지(adj-R2, Mallow-cp)를 사용하였다.

1) Adj-R2

```
 [1] 0.0063272469 0.0029120402 0.0017675774 0.0008251899 -0.0001120533 -0.0012673992 -0.0012751635
 [8] 0.0112507610 0.0080395941 0.0080386352 0.0067229220 0.0059583643 0.0054531624 0.0054126512
[15] 0.0150692099 0.0133358241 0.0117740130 0.0105264129 0.0103803377 0.0099879413 0.0097917048
[22] 0.0172811040 0.0152843972 0.0138505697 0.0138132156 0.0126788412 0.0122426684 0.0122276049
[29] 0.0162090333 0.0160812386 0.0160428221 0.0143170091 0.0140781384 0.0125961578 0.0116689508
[36] 0.0150407338 0.0149462004 0.0148387626 0.0131555255 0.0104558584 0.0101146078 0.0055351008
[43] 0.0137748622
```

분석 결과, 가장 높게 나온 값인 0.017로 22번째의 모델을 택한다. 독립변수가 6개인 모델로 변수를 선택하였다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 468.3511    21.4332  21.852  < 2e-16 ***
X3            6.1691     2.3384   2.638  0.00850 **
X4           -5.2325     1.7718  -2.953  0.00324 **
X5            0.4351     0.2619   1.661  0.09704 .
X7          -24.2601    11.9240  -2.035  0.04223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 317.1 on 781 degrees of freedom
Multiple R-squared:  0.02229,   Adjusted R-squared:  0.01728
F-statistic: 4.451 on 4 and 781 DF,  p-value: 0.00146
```

검정 결과, 5%의 경우, X5변수가 유의성을 띄지는 않지만, 1%를 기준으로 하면, 유의수준을 벗어나는 값이 존재하지는 않았다.

null_model 과 adj-R2 분석

```
Model 1: Y ~ X3 + X4 + X5 + X7
Model 2: Y ~ 1
  Res.Df      RSS Df Sum of Sq      F  Pr(>F)
1    781 78509820
2    785 80299584 -4  -1789764 4.4511 0.00146 **
```

비교한 결과, adj-R2의 모델이 유의한 것으로 나타났다.

2) Mallow-cp

```
> result_regfit$cp
 [1]  7.920484 10.635404 11.545195 12.294346 13.039408 13.957851 13.964023  5.003976  7.553438
[10]  7.554199  8.598792  9.205802  9.606900  9.639063  2.973683  4.348123  5.586518  6.575768
[19]  6.691595  7.002734  7.158335  2.223378  3.804587  4.940047  4.969628  5.867948  6.213357
[28]  6.225286  4.074828  4.175900  4.206283  5.571219  5.760140  6.932231  7.665554  6.000113
[37]  6.074783  6.159646  7.489202  9.621616  9.891164 13.508427  8.000000
```

분석 결과, mallow-cp 13.9640로 10번째 값이 가장 높았고, 10번째 모델을 선택하였다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 488.2238    18.4843  26.413   <2e-16 ***
X4           -4.1002     1.6800  -2.441   0.0149 *
X5            0.4032     0.2629   1.534   0.1255
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 318.5 on 783 degrees of freedom
Multiple R-squared:  0.01057,   Adjusted R-squared:  0.008039
F-statistic: 4.181 on 2 and 783 DF,  p-value: 0.01563
```

검정 결과, 0.5를 넘는 값은 존재하지 않았다. 두개의 변수 선택 과정 중 R-squared가 더 높은 adj-R2모델을 사용한다.

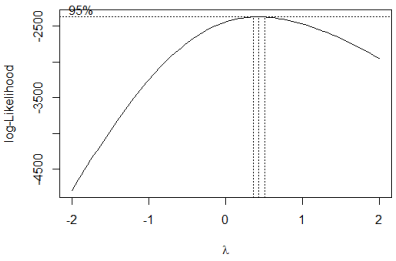
#단계별 회귀 검정

```
Coefficients:
(Intercept)           X4           X3           X7           X5
   468.3511      -5.2325       6.1691     -24.2601       0.4351
```

단계별 회귀를 실시하여 adj-R2에서 선택된 변수 4개 중 더 유의미한 설명변수가 있는지 확인하기 위해 단계별회귀를 실시했다. 실시한 결과, 변수 4개 모두 유의미해 모두 선택하였다.

#변수변환

#boxcox 변환



boxcox 결과, lambda=0.5로 가정한 후 변수변환을 진행한다.

Y

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 468.3511    21.4332  21.852  < 2e-16 ***
X3            6.1691     2.3384   2.638  0.00850 **
X4           -5.2325     1.7718  -2.953  0.00324 **
X5            0.4351     0.2619   1.661  0.09704 .
X7          -24.2601    11.9240  -2.035  0.04223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 317.1 on 781 degrees of freedom
Multiple R-squared:  0.02229,   Adjusted R-squared:  0.01728
F-statistic: 4.451 on 4 and 781 DF,  p-value: 0.00146
```

log

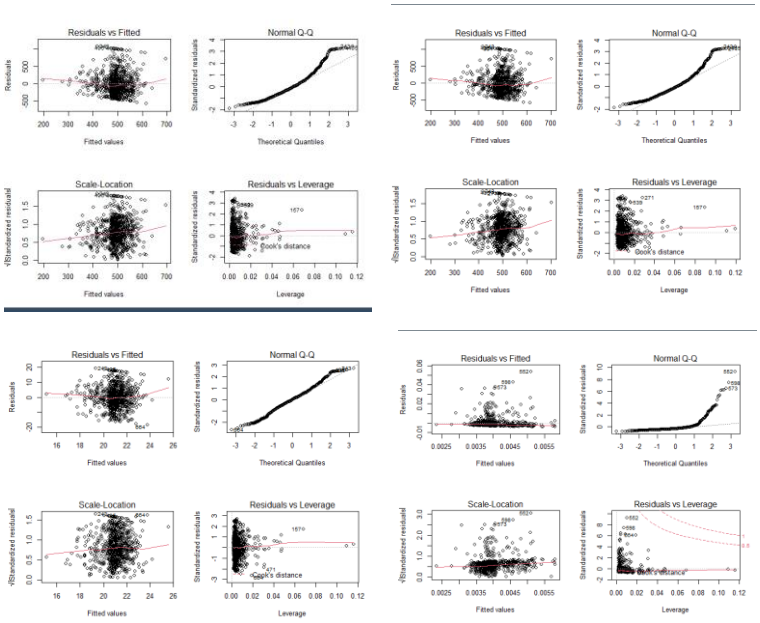
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.895329   0.054095 108.982   <2e-16 ***
X3           0.010653   0.005902   1.805   0.0715 .
X4          -0.011519   0.004472  -2.576   0.0102 *
X5           0.001077   0.000661   1.630   0.1035
X7          -0.036126   0.030094  -1.200   0.2303
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8002 on 781 degrees of freedom
Multiple R-squared:  0.01449,   Adjusted R-squared:  0.009444
F-statistic: 2.871 on 4 and 781 DF,  p-value: 0.02227
```

sqrt

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.434246   0.484772  42.152  < 2e-16 ***
X3           0.126701   0.052890   2.396  0.01683 *
X4          -0.114965   0.040073  -2.869  0.00423 **
X5           0.010353   0.005924   1.748  0.08090 .
X7          -0.465212   0.269693  -1.725  0.08493 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.171 on 781 degrees of freedom
Multiple R-squared:  0.01999,   Adjusted R-squared:  0.01498
F-statistic: 3.984 on 4 and 781 DF,  p-value: 0.003308
```

1/Y

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.100e-03  3.928e-04  10.438   <2e-16 ***
X3          -6.993e-06  4.286e-05  -0.163   0.8704
X4           5.846e-05  3.247e-05   1.801   0.0722 .
X5          -4.785e-06  4.800e-06  -0.997   0.3191
X7          -3.104e-05  2.185e-04  -0.142   0.8871
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.005811 on 781 degrees of freedom
Multiple R-squared:  0.005635,  Adjusted R-squared:  0.0005422
F-statistic: 1.106 on 4 and 781 DF,  p-value: 0.3523
```



변환 결과, sqrt 변수 변환을 했을 때, 정규성을 아주 잘 따르는 것을 확인할 수 있었다.

#Final model 회귀 모형의 선택

```
> sum(resid(Final_model)^2)
[1] 40159.72
> press
[1] 40864.26
```

```
> # R2 vs R2_predic
> 1-(SSE/SST) # R2 = 0.9994998
[1] 0.9994998
> 1-(press/SST) # R2_predic = 0.9994911
[1] 0.9994911
```

#SSE VS PRESS

40159.72 vs 40864.26으로 차이가 크지 않아 예측의 정확도가 높다고 할 수 있다.

R2 vs R2-predic

0.9994 VS 0.9995 값으로 차이가 거의 없어 예측의 정확도가 높다고 할 수 있고, 최종모형을 새로운 자료에 적용할 수 있다고 판단된다.

최종 모형: Y= 0.12670 X3 -0.11496 X4 + 0.01035 X5 -0.46521 X7

III 결론

1) 종속변수: 수면

```
> Final_model

Call:
lm(formula = log(Y) ~ X1 + X2 + X3 + X4 + X5 + X7)

Coefficients:
(Intercept)           X1           X2           X3           X4           X5           X7  
  2.6563564   -0.0133275    0.0039352   -0.0045339   -0.0021728    0.0001776   -0.0109004  
```

최종 모델:

- L_BR_FQ : 최근 1년동안 1주동안 아침식사 빈도(식사)/ BD2_14: 한 번에 마시는 음주량_잔(음주)/ BE8_1: 평소 하루 앉아서 보내는 시간(건강)/ BS3_2: 하루평균 흡연량(흡연)/ N_SUGAR: 1일 당 섭취량(식사 및 건강)/ N_CHO: 1일 탄수화물 섭취량(식사 및 건강)/ BE3_32: 걷기 지속 시간(건강)

2) 종속변수: 월평균 가구총소득

```
Call:
lm(formula = sqrt(Y) ~ X3 + X4 + X5 + X7)

Coefficients:
(Intercept)           X3           X4           X5           X7
   20.43425      0.12670     -0.11496      0.01035     -0.46521
```

최종 모델:

- BE8_1: 평소 하루 앉아서 보내는 시간(건강)/ BS3_2: 하루평균 흡연량(흡연)/ N_SUGAR: 1일 당 섭취량(식사 및 건강)/ BE3_32: 걷기 지속 시간(건강)

수면의 증가, 월평균 가구총소득의 증가= 행복의 증가라고 가정하였을 때, 1번째 회귀 최종 모델을 보면, 음주량, 1일 당 섭취량이 증가할수록 같이 증가하는 것을 알 수 있었다. 2번째 회귀 최종 모델을 보면, 앉아서 보내는 시간, 1일 당 섭취량이 증가할수록 같이 증가하는 것을 확인할 수 있었다. 이 두 가지의 모델의 공통점을 보자면, 흡연량이 적을수록, 1일 당 섭취량이 많을수록 Y와의 양의 상관관계를 갖는다. 즉, 흡연량이 적고, 1일 당 섭취량이 많을수록 행복도도 증가한다는 것을 이 연구를 통해 확인할 수 있었다. 물론 이 연구를 통해 행복의 정도를 단정지을 수는 없지만, 흡연을 과다하게 많이 하는 사람들에게는 흡연량을 줄이는 방안으로 권고할 수 있겠으며, 어느 정도의 당 섭취량은 있어야 인간의 행복에 영향을 미친다는 사실 또한 알리며 연구를 마치고자 한다.

한계 및 느낀점: '행복'을 주제로 정하다 보니, 변수를 선정하는데 있어서 상당히 추상적인 주제로 회귀 모델을 분석하는데 어려움을 겪었습니다. 수면의 증가, 월평균 가구총소득의 증가를 행복의 증가라고 가정을 한 상태로 분석을 진행했기에, 최종 모형이 행복과 100% 연관이 있다는 것이라고는 판단을 하기에 한계점이 있는 것 같습니다. 또한 아쉬웠던 점은, 결측치 제거로 인한 표본의 감소입니다. 표본의 감소를 막기 위해 '모름'을 0으로 바꾸려고 했으나, 변수변환의 과정에서 Y=0이 되어버리면 변환 자체가 불가하기에 제거를 택했습니다. 수면의 결측값이 적었다면 분석하는 것에 있어서 조금 더 명확한 결론이 나오지 않았을까 합니다. 반면, 의외였던 점은 걷는 시간(운동)이 증가할수록 행복의 정도도 증가할 것이라고 생각했으나, 두 모델 다 음의 상관관계를 갖는다는 점에서는 놀라웠습니다.