청소년의 스마트폰 중독에 관한 로지스틱 분석

1829044 통계학과

황지현

목차

서론

스마트폰은 우리 일상에서 없어서는 안될 필수재이다. 카메라, 인터넷, 전화, 문서업무, 여가에 활용하는 등 컴퓨터가 없어도 대부분의 일을 스마트폰에서 해결할 수 있을 정도로 활용도가 높다. 이처럼 스마트폰 기술이 고도화되고 사람들의 관심이 전통매체에서 스마트폰으로 옮겨지자 이를 활용하려는 애플리케이션들의 발달도 같이 진행되었다. 예를 들면, 작은 핸드폰에서도 고화질로 영화를 감상할 수 있는 넷플릭스, 인기콘텐츠에 광고를 붙여 수익을 창출하는 유튜브, 사람사이의 소통을 온라인으로 대체해주는 페이스북, 인스타그램 등 매력적인 애플리케이션이 많이 생겼다.

이런 애플리케이션들은 우리의 삶의 재미와 질을 높여주는 긍정적인 기능을 갖고 있지만, 한편으로는 부정적인 영향을 끼치기도 한다. 스마트폰 자체에 쏟는 시간이 많아지면서 스마트폰 중독현상이 그 예이다. 통계청의 *스마트폰 과의존실태조사 보고서*에 따르면 17년도에 조사된 스마트폰 과의존위험군은 18.6 %로 이전부터 상승하고 있다고 한다. 특히 청소년들에게서 그 증가폭이 크다.

이러한 부작용을 막기 위해서 우리는 청소년들이 어떤 환경에서 더 중독 위험이 높아질 수 있는지 파악해야 한다. 그래서 필자는 통계청의 스마트폰과의존실태조사 자료를 이용하여 청소년이 스마트폰에 중독될 확률이 어떤 환경에서 높아지는지 로지스틱 모형을 이용하여 분석하려한다.

자료설명

자료출처: MDIS https://mdis.kostat.go.kr/

(자료이용>다운로드 서비스>정보통신/과학기술>스마트폰 과의존 실태조사)

<스마트폰 과의존 실태조사> 자료는 스마트폰 과의존 예방·해소 정책의 성과 평가와 효과적인 정책 수립 및 개선방안도출을 위한 기초 정책통계 생산을 목적으로 조사된 자료이고 전국 10,000개 가구 내 만3~69세 스마트폰(인터넷) 이용자를 대상으로 조사되었다.
추출한 자료는 2017,2018년에 수집된 자료이고, 자료개수는 29712, 변수개수는 12개이다.

변수설명 (추출한 변수중에 분석에 사용된 변수만 제시하였음)

1. 스마트폰 과의존위험여부(is_toxic)

0	1
아니다	그렇다

2. 월평균 가구소득($income\_per\_month$) (만원단위)

1	2	3	4	5	6
200미만	200이상 400미만	400이상 600미만	600이상 800미만	800이상 1000미만	1000이상

3. 맞벌이여부($working\_parent$)

0	1
그렇다	아니다

4. 성별(sex)

1	2
남성	여성

5.학령별($school\_age$)

2	3	4	5
초등학생	중학생	고등학생	대학생

분석

먼저 로지스틱 분석에 사용할 설명변수는 월평균 가구소득($income\_per\_month$), 맞벌이 여부($working\_parent$), 성별(sex), 학령별($school\_age$)이고 반응변수는 스마트폰 과의존위험여부(is_toxic)이다. $income_per_month$는 순서형 변수, $working\_parent$, sex는 이항변수, $school_age$는 명목형 변수로 설정하였다.

$is_toxic = 1$일 확률에 가장 적합한 모형을 찾기 위하여 계획한 단계는 다음과 같다.

1단계: 영모형, 한 개의 설명변수만 포함한 모형 간의 차이를 통해 유의한 변수를 뽑는다.

2단계: 후진제거법을 이용하여 모형을 단순화시킨다.

3단계: 1단계에 포함되지 않았으나 2단계이후에 유의할 가능성이 있는 변수를 추가한다.

4단계: 3단계 이후 모형에 남아있는 변수들 간에 상호작용이 존재하는지 검정한다.

5단계: 선택한 모형에 AIC를 적용한다.

6단계: 적합결여를 확인한다. (후속진단)

모든 단계에서 유의수준은 0.05로 고정한다.

- **1단계**

***-income_per_month* 변수의 유의성**

```
##                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                              5087     6128.4
## income_per_month  1   143.74      5086     5984.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value가 0.05보다 작으므로 $income_per_month$가 들어있는 모형이 채택된다.

$income_per_month$는 유의하다.

-$working\_parent$ 변수의 유의성

```
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           5087     6128.4
## working_parent  1   11.758      5086     6116.6 0.0006057 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value가 0.05보다 작으므로 $working_parent$가 들어있는 모형이 채택된다.

$working_parent$는 유의하다.

-sex 변수의 유의성

```
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                 5087     6128.4
## sex   1   3.1387      5086     6125.2  0.07645 .
```

p-value가 0.05보다 크므로 영모형이 채택된다.

sex는 유의하지 않다.

-$school\_age$ 변수의 유의성

$school_age$는 명목형 변수이지만, 순서형 변수로 할당하여 학령이 로짓에 대해 선형효과가 있다고 가정할 수 있다. 이때 순서형 변수로 만든 모형이 명목형 변수로 만든 모형보다 더 간단하기 때문에 만약 더 잘 적합된다면 $school_age$는 순서형 변수로 생각해야한다.

```
## Analysis of Deviance Table
##
## Model 1: is_toxic ~ school_age
## Model 2: is_toxic ~ factor(school_age)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5086     6115.6
## 2      5084     6062.7  2   52.849 3.341e-12 ***
```

결과를 보면 더 복잡한 모형, 즉 명목형 변수로 만든 모형이 유의하다는 결과를 얻는다. 따라서 $school_age$는 명목형변수로 생각해야한다.

따라서 선택된 변수로 만들어진 모형은 다음과 같다.

$$is\_toxic \sim income\_per\_month + working\_parent + factor(school\_age)$$

- 2단계

-$income_per_month$가 포함된 모형과 포함되지 않은 모형의 비교

```
## Analysis of Deviance Table
##
## Model 1: is_toxic ~ working_parent + factor(school_age)
## Model 2: is_toxic ~ income_per_month + working_parent + factor(school_age)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5083     6050.9
## 2      5082     5851.5  1   199.44 < 2.2e-16 ***
```

p-value가 0.05보다 작으므로 $income_per_month$가 포함된 모형이 채택된다.

-$working_parent$가 포함된 모형과 포함되지 않은 모형의 비교

```
## Analysis of Deviance Table
##
## Model 1: is_toxic ~ income_per_month + factor(school_age)
## Model 2: is_toxic ~ income_per_month + working_parent + factor(school_age)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5083     5907.3
## 2      5082     5851.5  1   55.798 8.03e-14 ***
```

p-value가 0.05보다 작으므로 $working_parent$가 포함된 모형이 채택된다.

-$factor(school_age)$가 포함된 모형과 포함되지 않은 모형의 비교

```
## Analysis of Deviance Table
##
## Model 1: is_toxic ~ income_per_month + working_parent
## Model 2: is_toxic ~ income_per_month + working_parent + factor(school_age)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5085     5931.5
## 2      5082     5851.5  3   79.966 < 2.2e-16 ***
```

p-value가 0.05보다 작으므로 $factor(school_age)$가 포함된 모형이 채택된다.

- 3단계

1단계에서 유의하지 않았던 sex를 다시 모형에 넣어 기존 모형과 비교한다.

```
## Analysis of Deviance Table
##
## Model 1: is_toxic ~ income_per_month + working_parent +
factor(school_age)
## Model 2: is_toxic ~ income_per_month + working_parent +
factor(school_age) + sex
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5082     5851.5
## 2      5081     5849.0  1   2.5389   0.1111
```

p-value가 0.1111이므로 복잡한 모형인 sex를 포함한 모형이 기각된다.

- 4단계

교호작용 항이 유의한지 알아보기 위하여 세개의 교호작용 항을 순차적으로 넣어 모형을 비교해 본다.

- $income_per_month * working_parent$을 넣었을 때

```
## Analysis of Deviance Table
##
## Model 1: is_toxic ~ income_per_month + working_parent + factor(school_age)
## Model 2: is_toxic ~ income_per_month + working_parent + factor(school_age)
##     +income_per_month * working_parent
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5082     5851.5
## 2      5081     5834.8  1    16.72 4.333e-05 ***
```

p-value가 0.05보다 작으므로 교호작용을 포함한 모형이 채택된다.

- $income_per_month * factor(school_age)$을 넣었을 때

```
## Analysis of Deviance Table
##
## Model 1: is_toxic ~ income_per_month + working_parent + factor(school_age)
+
##     income_per_month * working_parent
## Model 2: is_toxic ~ income_per_month + working_parent + factor(school_age)
+
##            income_per_month   *   working_parent   +   income_per_month   *
factor(school_age)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5081     5834.8
## 2      5078     5812.4  3   22.352 5.51e-05 ***
```

p-value가 0.05보다 작으므로 교호작용이 포함된 모형이 채택된다.

- $working_parent * factor(school_age)$를 넣었을 때

```
## Analysis of Deviance Table
##
## Model 1: is_toxic ~ income_per_month + working_parent +
factor(school_age) +
##     income_per_month * working_parent + income_per_month *
factor(school_age)
## Model 2: is_toxic ~ income_per_month + working_parent +
factor(school_age) +
##     income_per_month * working_parent + income_per_month *
factor(school_age) +
##     working_parent * factor(school_age)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5078     5812.4
## 2      5075     5807.8  3    4.634   0.2006
```

p-value가 0.05보다 크므로 $working_parent * factor(school_age)$을 포함한 모형은 기각된다.

따라서 최종적으로 선택된 모형은 다음과 같다.

$$is\_toxic \sim income\_per\_month + working\_parent + factor(school\_age) + income\_per\_month * woiking\_parent + income\_per\_month * factor(school\_age)$$

- 5단계

위의 모형에 AIC를 적용한 결과는 다음과 같다.

```
## Start:  AIC=5832.44
## is_toxic ~ income_per_month + working_parent + factor(school_age) +
##     income_per_month * working_parent + income_per_month *
factor(school_age)
##
##                                      Df Deviance    AIC
## <none>                                    5812.4 5832.4
## - income_per_month:working_parent     1   5829.9 5847.9
## - income_per_month:factor(school_age)  3   5834.8 5848.8

##
## Call:  glm(formula = is_toxic ~ income_per_month + working_parent +
##     factor(school_age) + income_per_month * working_parent +
##     income_per_month * factor(school_age), family = binomial,
##     data = teen_user)
##
## Coefficients:
##                         (Intercept)                    income_per_month
##                             1.58008                             -0.63145
##                      working_parent                  factor(school_age)3
```

```
##                             -1.26839                              1.32481
##                 factor(school_age)4                 factor(school_age)5
##                             0.42852                              0.34506
##     income_per_month:working_parent
income_per_month:factor(school_age)3
##                             0.24962                             -0.25046
## income_per_month:factor(school_age)4
income_per_month:factor(school_age)5
##                            -0.09866                              0.15914
##
## Degrees of Freedom: 5087 Total (i.e. Null);  5078 Residual
## Null Deviance:       6128
## Residual Deviance: 5812  AIC: 5832
```

최종선택된 모형은 기존의 선택한 모형과 같다.

- 6단계

적합결여를 보기위하여 그룹화되지 않은 자료를 3개의 변수에 대해서 그룹화한 자료로 만들고 각 수준에 대해서 표준화 잔차를 구하였다. 다음은 10개의 자료만 나타낸 것이다.

```
## # A tibble: 91 x 7
## # Groups:   income_per_month, working_parent, school_age, is_toxic [91]
##    income_per_month working_parentschool_age is_toxic    n fitted
std.res
##              <int>         <int>      <int>    <dbl> <int>  <dbl>   <dbl>
##  1               1             1          2        0     3  0.362 -0.787
##  2               1             1          3        0     1  0.387 -0.829
##  3               1             1          4        0     2  0.387 -0.829
##  4               1             1          5        0     1  0.387 -0.829
##  5               1             2          2        0    10  0.399 -0.852
##  6               1             2          3        0    10  0.425 -0.896
##  7               1             2          3        1     7  0.425  1.21
##  8               1             2          4        0    27  0.425 -0.896
##  9               1             2          4        1    10  0.425  1.21
## 10               1             2          5        0     7  0.425 -0.896
## # ... with 81 more rows
```

다음은 표준화 잔차값만 나열한 것이다.

```
data$std.res

##         1         2         3         4         5         6         7
## -0.7871982 -0.8291911 -0.8291911 -0.8291911 -0.8515861 -0.8958797
1.2117575
##         8         9        10        11        12        13        14
## -0.8958797  1.2117575 -0.8958797  1.2117575 -0.8249641  1.3008604 -
```

```
0.8690414
##         15         16         17         18         19         20         21
##  1.2325566 -0.8690414  1.2325566 -0.8690414  1.2325566 -0.8925741
1.2010223
##         22         23         24         25         26         27         28
## -0.9391803  1.1366494 -0.9391803  1.1366494 -0.9391803  1.1366494 -
0.8677775
##         29         30         31         32         33         34         35
##  1.2243781 -0.9143734  1.1603834 -0.9143734  1.1603834 -0.9143734
1.1603834
##         36         37         38         39         40         41         42
## -0.9392717  1.1308619 -0.9886350  1.0705947 -0.9886350  1.0705947 -
0.9886350
##         43         44         45         46         47         48         49
##  1.0705947 -0.9164501  1.1569845 -0.9660070  1.0969069 -0.9660070
1.0969069
##         50         51         52         53         54         55         56
## -0.9660070  1.0969069 -0.9925007  1.0692042 -1.0450587  1.0126079 -
1.0450587
##         57         58         59         60         61         62         63
##  1.0126079 -1.0450587  1.0126079 -0.9718993  1.0978715 -1.0248617
1.0412771
##         64         65         66         67         68         69         70
## -1.0248617  1.0412771 -1.0248617  1.0412771 -1.0531790  1.0151812 -
1.1093545
##         71         72         73         74         75         76         77
##  0.9617951 -1.1093545  0.9617951 -1.1093545  0.9617951 -1.0351683
1.0462917
##         78         79         80         81         82         83         84
## -1.0919726  0.9927144 -1.0919726  0.9927144 -1.0919726  0.9927144 -
1.1223349
##         85         86         87         88         89         90         91
##  0.9680001 -1.1825230  0.9173463 -1.1825230  0.9173463 -1.1825230
0.9173463
```

여기서 변수를 조합하면 수준 수가 6*2*2*4=96개가 나와야 하는데 91개의 수준밖에 없는 이유는 96개 중에 자료가 없는 수준이 5개이기 때문이다.

표준화잔차중 절대값이 2보다 큰 것이 있는지 확인하였다.

```
data$std.res[abs(data$std.res)>2]
```

```
## named numeric(0)
```

모두 절대값이 2보다 작아 모형이 잘 적합되었다고 생각된다.

- **모형**

적합된 로지스틱 선형식은 다음과 같다.

$x =$
$(\ income\_per\_month, working\_parent, school\_age3, school\_age4, school\_age5)$

$$\pi(x) = P(is\_toxic = 1)$$

$$\begin{aligned} logit[\pi(x)] = {} & 1.58 - 0.63 * income\_per\_month - 1.2 * working\_parent \\ & + 1.32 * school\_age3 + 0.42 * school\_age4 + 0.35 \\ & * \text{school\_age5} + 0.25 * income\_per\_month * working\_parent \\ & - 0.25 * income\_per\_month * school\_age3 - 0.1 \\ & * income\_per\_month * school\_age4 + 0.16 \\ & * income\_per\_month * school\_age5 \end{aligned}$$

$$School\_age3 = \begin{cases} 1, school\_age = 3 \\ 0, o.w \end{cases}$$

$$School\_age4 = \begin{cases} 1, school\_age = 4 \\ 0, o.w \end{cases}$$

$$School\_age3 = \begin{cases} 1, school\_age = 5 \\ 0, o.w \end{cases}$$

결론

-$income_per_month$의 효과

$working_parent = 1, school_age = 4$로 고정시켰을 때 새로 얻어지는 선형식은 다음과 같다.

$$logit[\pi(x)] = 0.8 - 0.48 * \text{income\_per\_month}$$

$Income_per_month$가 1씩 증가할 때 중독일 확률에 대한 오즈는 $e^{-0.48}$=0.62배가 된다. 즉, 외벌이 가정이고 고등학생일때, 가구소득이 한 단위 증가하면 스마트폰 중독일 확률은 점점 작아진다.

$Working_parent = 0, school_age = 4$로 고정시켰을 때 새로 얻어지는 선형식은 다음과 같다.

$$logit[\pi(x)] = 2 - 0.73 * income\_per\_month$$

$Income_per_month$가 1씩 증가할 때 중독일 확률에 대한 오즈는 $e^{-0.73} = 0.48$배가 된다.

즉, 맞벌이 가정이고 고등학생일 때, 가구소득이 한 단위 증가하면 스마트폰 중독일 확률은 점점 작아진다.

이 때, 고등학생의 경우, 가구소득이 증가할 수록 맞벌이 가정이 외벌이 가정보다 스마트폰 중독일 확률이 더 빠르게 감소한다. 이는 $income_per_month$와 $working_parent$의 교호작용이 존재하여 나타난 결과이다.

-$working_parent$의 효과

$Income_per_month = 1, school_age = 4$로 고정시켰을 때 새로 얻어지는 선형식은 다음과 같다.

$$logit[\pi(x)] = 1.27 - 0.95 * working\_parent$$

$Working_parent$가 1에서 0로 바뀔 때 중독일 확률에 대한 오즈는 $e^{0.95}$=2.59배가 된다. 즉 가구소득이 200만원 미만이고 고등학생일 때, 스마트폰 중독일 확률에 대한 오즈는 맞벌이일때가 외벌이 일때보다 2배이상 높다.

$Income_per_month = 1, school_age = 3$으로 고정시켰을 때 새로 얻어지는 선형식은 다음과 같다.

$$logit[\pi(x)] = 2.02 - 0.95 * working\_parent$$

$Working_parent$가 1에서 0으로 바뀔 때 중독일 확률에 대한 오즈는 $e^{0.95}$=2.59배가 된다. 이는 위의 모형과 결과가 같은데,$working_parent$와 $school_age$간의 교호작용이 없기 때문에 나온 결과이다. 하지만 어떤 중학생, 고등학생에 대하여 두 가정 모두 가구소득이 200만원 미만일 때 부모님 맞벌이 여부와는 관계없이 스마트폰 중독일 확률이 중학생이 고등학생보다 $e^{0.75} = 2.12$배가 됨을 알 수 있다.

-$school_age$의 효과

$Income_per_month = 1, working_parent = 1$로 고정시켰을 때 새로 얻어지는 선형식은 다음과 같다.

$$logit[\pi(x)] = \text{-}0.25 + 1.07\text{*school\_age3} + 0.32\text{*school\_age4} + 0.51\text{*school\_age5}$$

$school_age$가 2에서 3으로 바뀔 때 중독일 확률에 대한 오즈는 $e^{1.07} = 2.92$배가 된다. 즉 가구소득이 200만원 미만이고 외벌이 가정일 때, 스마트폰 중독일 확률에 대한 오즈는 중학생이 초등학생의 약 3배이다.

$school_age$가 3에서 4로 바뀔 때 중독일 확률에 대한 오즈는 $e^{-0.75} = 0.47$배가 된다. 즉 가구소득이 200만원 미만이고 외벌이 가정일 때 스마트폰 중독일 확률에 대한 오즈는 고등학생이 중학생보다 약 0.5배 낮다. 중학생보다 고등학생이 스마트폰에 중독될 확률이 낮다.

$school_age$가 4에서 5로 바뀔 때 중독일 확률에 대한 오즈는 $e^{0.19} = 1.21$배가 된다. 즉 가구소득이 200만원 미만이고 외벌이 가정일 때, 스마트폰 중독일 확률에 대한 오즈는 대학생이 고등학생의 1.21배이다.

Income_per_month=3,working_parent=1로 고정시켰을 때 새로 얻어지는 선형식은 다음과 같다.

$$logit[\pi(x)] = -0.76 + 0.57 * school\_age3 + 0.12 * school\_age4 + 0.83 * school\_age5$$

School_age가 2에서 3으로 바뀔 때 중독일 확률에 대한 오즈는 $e^{0.57}$=1.77배가 된다. 즉 가구소득이 400만원 이상 600만원 미만이고 외벌이인 가정에서 스마트폰 중독일 확률에 대한 오즈는 중학생이 초등학생보다 0.77배 높다.

School_age가 3에서 4로 바뀔 떄 중독일 확률에 대한 오즈는 $e^{-0.45} = 0.64$배가 된다. 즉 가구소득이 400만원 이상 600만원 미만이고 외벌이이 가정에서 스마트폰 중독일 확률에 대한 오즈는 중학생이 초등학생보다 0.36배 낮다.

School_age가 4에서 5로 바뀔 때 중독일 확률에 대한 오즈는 $e^{0.71} = 2.03$배가 된다. 즉 가구소득이 400만원 이상 600만원 미만이고 외벌이인 가정에서 스마트폰 중독일 확률에 대한 오즈는 대학생이 고등학생의 약 2배이다.

Income_per_month와 school_age는 교호작용이 있기 때문에 income_per_month를 다르게 고정시킨 두 모형에 대해서 측정한 대응되는 두 오즈값은 다를 수밖에 없다. 위의 분석에서 확인할 수 있다.