2021-2학기 회귀분석(1139301-01)

Final project

기말프로젝트 연구계획서

- 국민건강영양조사 자료를 이용하여 -

<성인근로자의 근무형태 및 수면시간이 건강 관련 삶의 지수에 미치는 영향>

학번 : 20160393

이름 : 황건하

I 서론

요약

자료는 국민건강영양조사 제8기 1차년도(2019년)자료를 이용하였다. 모름, 무응답은 결측치로 간주했고, 결측치는 모두 제거하고 분석하였다. 대상을 야간근무자와 주간근무자로 선정하였으며 삶의 질의 차이와 흡연, 음주, 운동이 교대 근무시 삶의 질과 건강상태에 미치는 관련성을 확인하였다. 본 보고서는 야간근무자에 대한 삶의 질 향상에 건강관리 기초자료를 제공하는 것에 의의를 둔다.

1. 연구 배경 및 이유

국민생활과 이용자들의 편의를 위한 공공서비스 증가로 노동시간이 1 일 24 시간으로 확대된 상태이다. 병원 물류업, 공장, 편의점, 식당, 경비와 보안, 교통, 청소 및 미화 등 윤택하고 편리한 현대 사회가 유지되는 데는 밤에 일하는 수많은 사람들이 기여한다. 이렇듯 야간작업, 교대근무는 현대사회에서 필수적인 형태의 노동이다. 우리나라에서는 교대근무에 대한 법적 정의는 없으나 대체로 한 근무조에서 다른 근무조로 업무가 인계될 수 있는 업무 활동을 의미한다. 보통 야간근무는 평소에 행동 및 수면패턴과 다르게 움직이기 때문에 생체리듬을 파괴하여 여러 신체적, 정신적 건강문제를 야기한다고 알려져 있다. 필자 또한 최근부터 주 2 일 편의점 야간 아르바이트를 시작하였는데 밤낮이 바뀐 올빼미형 인간이 되어 생활에 적응하기 어렵다. 이러한 불규칙한 생활양식은 식습관, 운동 등을 비롯한 생활양식 뿐만 아니라 고혈압, 당뇨병 등 건강문제와 우울증과 같은 정신적인 문제에 영향을 끼쳐 사회적·심리적 문제를 초래한다. 교대근무가 미칠 수 있는 삶의 질과 만성질환에 대한 지속적인 연구가 필요하며, 국민들의 질환을 조기에 발견하고 관리하는 것이 중요하다고 생각한다.

2. 연구 목표

야간근무자와 주간근무자의 삶의 질의 차이를 확인하고 수면 관리의 중요성을 알아보고자 근무형태 및 수면시간이 건강 관련 삶의 지수에 미치는 영향에 대해 알아보고자 한다. 이 프로젝트를 통해 정신적, 육체적 및 수면 관리의 중요성을 통하여 근로자들의 건강관리 기초자료를 제공하는 것이 목표이다.

II 본론

연구내용

데이터 전처리를 위해 R 프로그램의 dplyr패키지를 활용하였다. 주로 filter, select, replace 함수를 이용하였고, 중선형회귀모형으로 만들었다. 먼저 데이터 추출을 한 뒤, 변수 선택 기준에 맞춰서 변수를 선정하고, 회귀진단 후 최종 모형을 확인하였다. 연구대상은 최초 표본수가 8,110명이었는데, 19세 미만 대상자와 결측 데이터를 제거하고, 설문지 데이터가 누락된 개인도 데이터에서 제외하였다. 그래서 표본수는 142명(남자: 130, 여자: 12)이었는데 최종모형에서의 표본수는 135(남자: 123, 여자: 12)명으로 추정되었다. 이번 프로젝트에서 야간근무자와 주간근무자와의 삶의 지수와 수면상태에 대한 차이를 분석하고 야간근무자를 위한 근로환경 개선에 집중하고자 한다.

1. 변수의 선정 및 정의

1) 종속변수

종속변수는 건강 관련 삶의 질 지수인 HINT-8 이다. HINT-8 은 우리나라 사람들의 HRQOL 을 측정하기 위해 질적 및 양적 연구 방법을 사용하여 개발된 도구로서 4 개의 건강영역에 기반한 8 개의 항목으로 이루어져 있다. 각 항목은 문제의 정도에 따라 4 수준으로 되어 구성되었는데 아무 문제가 없는 경우 수준 1 이고 문제가 심각한 경우 수준 4 이다. HINT-8 으로 표현할 수 있는 건강상태의 수는 65,536 개(=4^8)로서 8 자리 숫자(8-digit profile)로 표현할 수 있다. 총 8 가지의 항목 CL(계단 오르기), PA(통증), VI(활력), WO(일하기), DE(우울), ME(기억하기), SL(수면), HA(행복) 등이 있다. 수준이 2,3,4 인 경우 1 로 환산되고, 나머지 경우 0 으로 환산이 된다. 이때 건강상태의 질 가중치 값(QW)을 이용하여 종속변수를 다음과 같이 하였다. 'y=1-QW'모형에서 상수항은 '11111111'의 상태에서의 불효용을 나타내는데 이를 종속변수로 이용한 모형을 탐색하였다.

HINT-8 지수 산출식 및 설명

HINT-8 지수 산출식

HINT-8 지수 = 1-(0.073 +	+ 0.018 x CL2 + 0.072 x CL3 + 0.122 x CL4 + 0.055 x PA2 + 0.116 x PA3 + 0.188 x PA4 + 0.019 x VI23 + 0.070 x VI4 + 0.004 x WO2 + 0.028 x WO3 + 0.036 x WO4 + 0.012 x DE2 + 0.044 x DE3 + 0.098 x DE4 + 0.014 x ME2 + 0.058 x ME3 + 0.109 x ME4 + 0.020 x SL3 + 0.090 x SL4 + 0.014 x HA2 + 0.068 x HA3 + 0.082 x HA4)

산출식 설명

항목	항목별 수준
CL: 계단 오르기	2: 수준 2 인 경우 1, 나머지 경우 0
PA: 통증	3: 수준 3 인 경우 1, 나머지 경우 0
VI: 활력	4: 수준 4 인 경우 1, 나머지 경우 0
WO: 일하기	(활력 항목에서 수준 2 또는 3 인 경우 VI23=1 이고
DE: 우울	나머지 경우 0 이 됨)
ME: 기억하기	
SL: 수면	
HA: 행복	

HINT-8 으로 표현할 수 있는 건강상태 중 가장 좋은 상태는 11111111 이고, 가장 나쁜 상태는 44444444 이다. 완전한 건강상태인 11111111 의 경우 HINT-8 지수는 1 이 되고 44444444 인 경우 위의 식에 따라 산출하면 0.132 가 된다. 즉 HINT-8 지수의 값은 0.132 ~ 1 의 범위를 갖는다. 1 에 가까운 값을 가질수록 좋은 건강상태를 의미하고 값이 작아질수록 나쁜 건강상태를 의미한다.

2) 독립변수

독립변수는 주중 하루 평균 수면시간, 주당 평균 근로시간, 걷기 지속시간, 하루평균 흡연량, 주당 평균 근로시간, 여가_중강도 신체활동 시간, 성별, 교육수준, 변형근로시간 등을 선정하였습니다. 선정 기준은 필자의 직감에 의해 선정하였다 사회인구학적 특성으로 성별을 선택하였고, 성별은 1이면 남성, 0이면 여성으로 분류하였다. 교육수준은 대

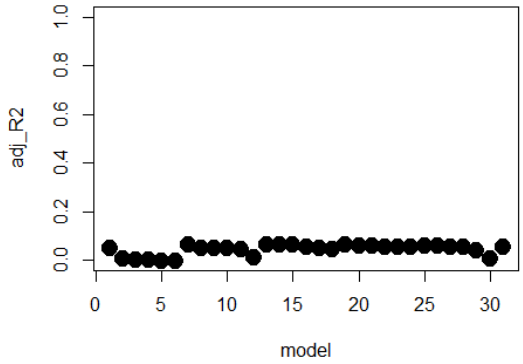
학졸업이면 0, 대학졸업이 아닐 경우 1로 분류하였다. 변형근로시간은 야간근무를 할 경우 0, 주간 또는 저녁근무를 할 경우 1로 분류하였다.

2. 통계 분석

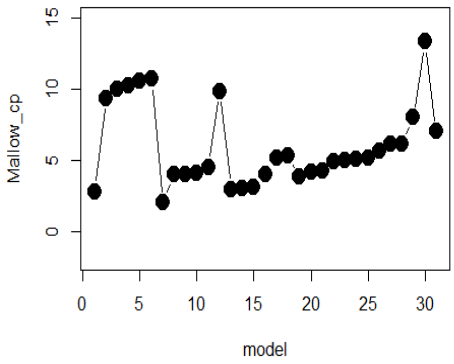
국민건강영양조사 자료는 복합표본설계방법을 사용하였으므로 층화, 집락, 가중치의 복합표본설계요소를 반영하여 복합표본 자료분석을 실시하였다. 분석은 탐색적 분석과 회귀모형으로 나누어 분석을 진행하였다. 회귀분석 전, 변수 간의 선정기준 4가지에 따라서 변수를 선택하였고, 탐색적 분석의 내용을 바탕으로 성별, 교육수준, 변형근로시간 여부를 독립변수로 하여 건강형태, 정신건강과 삶의 질과의 연관성을 알아보기 위하여 계단 오르기, 통증, 기운, 일하기, 우울, 기억, 잠자기, 행복을 보정한 후 중선형회귀분석을 시행하였다. 통계 분석은 R프로그램을 이용하였고, p-value<0.05인 경우 통계적으로 유의하다고 판단하였다.

연구결과

1. 변수선택 기준- 선택방법

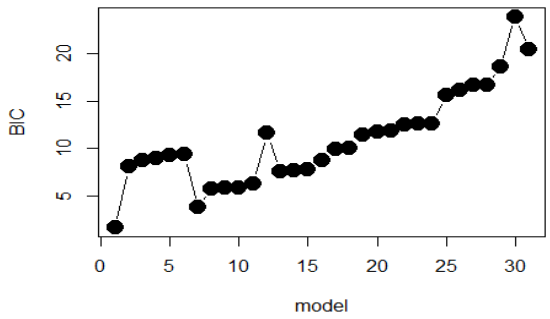


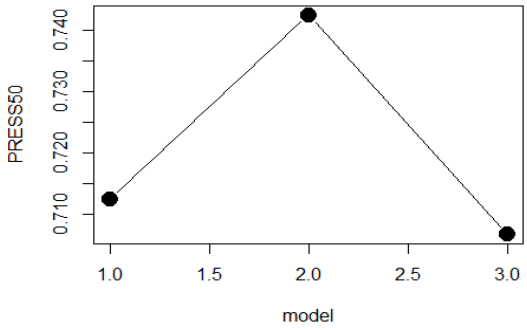
(1) 수정된 결정계수와 MSE



(2) Mallows-Cp

부분F검정을 한 결과 p-value가 0.05보다 작으므로 귀무가설을 기각할 수 있다. (1) 수정된 결정계수와 MSE 기준에서 7번째 모형에서 가장 큰 값을 도출할 수 있었다. 또한 (2) Mallows-Cp 기준에서 7번째 모형에서 가장 작은 값을 도출할 수 있었다. 이때 변수는 주중 하루 평균 수면시간과 여가_중강도 신체활동 시간이었다. 즉, 이 두 변수는 모형에 적합하다는 것을 의미한다.





(3) BIC (4) PRESS_p

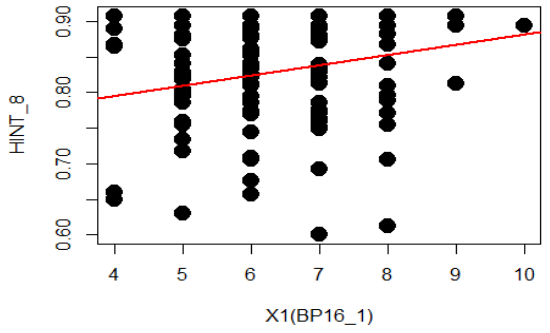
부분F검정을 한 결과 p-value가 0.05보다 작으므로 귀무가설을 기각할 수 있다. (3) BIC 기준에서 1번째 모형에서 가장 작은 값을 도출할 수 있었다. 또한 (2) PRESS_p 기준에서 변수가 모두 포함한 모형에서 가장 작은 값을 도출할 수 있었다. 이때 변수는 마찬가지로 주중 하루 평균 수면시간과 여가_중강도 신체활동 시간이었다. 즉, 이 두 변수는 모형에 적합하다는 것을 의미한다.

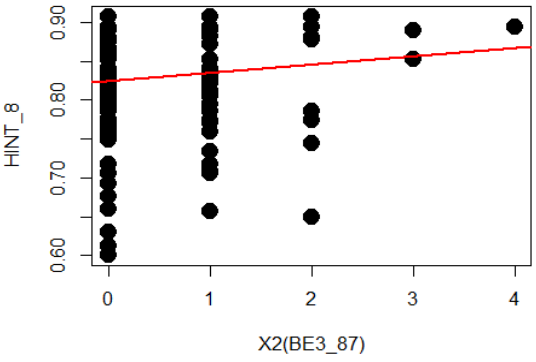
2. 변수선택 기준– 단계별 회귀, 전진선택법, 후진제거법

```
lm(formula = HINT_8 ~ BP16_1 + BE3_87, data = health_data)

Coefficients:
(Intercept)       BP16_1       BE3_87
    0.72635      0.01506      0.01282
```

세 기준 모두 동일한 결과가 나왔고, 변수를 2개로 선정하는 것이 맞다고 판단된다.

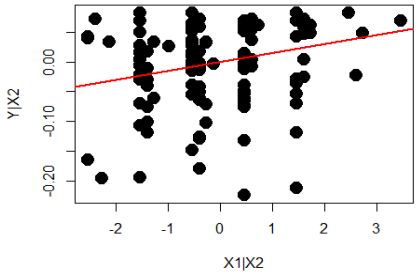
3. 1. 회귀진단· 모형확인 – 편회귀그림





(1) X1 만을 포함한 모형의 적합 (2) X2 만을 포함한 모형의 적합

(3.) X1 에 대한 편회귀그림




```
Call:
lm(formula = y.x2 ~ x1.x2)

Coefficients:
(Intercept)        x1.x2
 -4.575e-18    1.506e-02
```

편회귀그림이란 다른 설명변수들의 영향이 제거된 Y 와 Xi 의 순수한 관계를 보여주는 그림이다. X1(주중수면시간)에 대한 X1 에 대한 회귀 기울기는 1.5606e002 이다. 즉, HINT_8 지수와 주중수면시간은 선형적인 관계를 보이고 둘은 양의 상관관계를 보인다. 즉, 수면시간에 따라 HINT_8 지수에 영향을 준다는 것을 의미한다.

3. 2. 회귀진단· 모형확인 – 적합결여검정

```
Model 1: Y ~ X1
Model 2: Y ~ factor(X1)
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1    140 0.69117
2    135 0.68999  5 0.0011767 0.046 0.9987
```

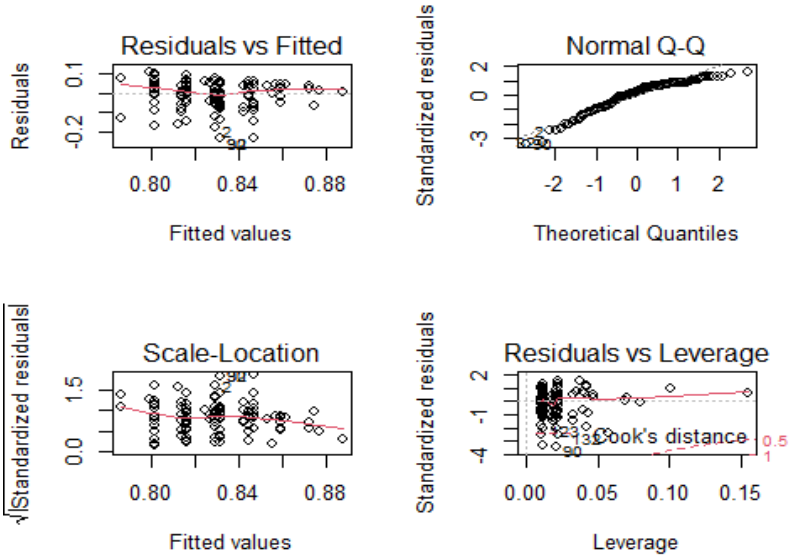
가설:H0: 가정된 모형 E[Yi] = B0+β1X1i VS H1: 가정된 모형은 옳지 않다

```
Model 1: Y ~ X1 + X2
Model 2: Y ~ factor(X1) + factor(X2)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    139 0.67761
2    131 0.67184  8 0.0057668 0.1406 0.9972
```

가설:H0: 가정된 모형 E[Yi] = B0+β1X1i+β2X2i VS H1: 가정된 모형은 옳지 않다

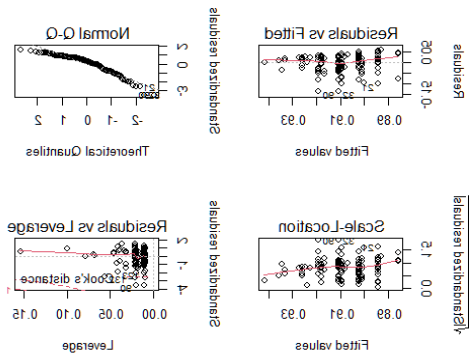
p-value가 0.05보다 작으므로 HINT_8지수에 대한 주중수면시간과 여가 중강도 신체활동에 대한 hint8지수가 있는 모형이 적절하다고 판단할 수 있다.

3. 3. 회귀진단· 모형확인 – 변수변환

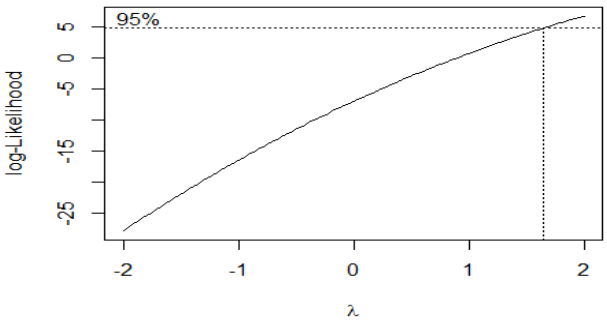


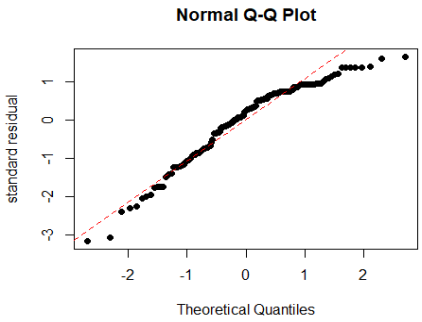

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.726351   0.032935  22.054  < 2e-16 ***
X1          0.015063   0.004917   3.063  0.00263 **
X2          0.012821   0.007688   1.668  0.09763 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06982 on 139 degrees of freedom
Multiple R-squared:  0.07541,   Adjusted R-squared:  0.06211
F-statistic: 5.668 on 2 and 139 DF,  p-value: 0.0043
```

잔차분석한 결과 잔차는 중심을 기준으로 잘 흩어져 있고, Q_Q도 선형적이라는 것을 봤을 때 굳이 변수변환을 할 필요가 없어보인다. 하지만 혹시나 하는 마음에 √Y 변환과 Box-Cox 변환을 해보았다.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.852857   0.018572  45.922  < 2e-16 ***
X1          0.008324   0.002773   3.002  0.00318 **
X2          0.007097   0.004335   1.637  0.10389
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03937 on 139 degrees of freedom
Multiple R-squared:  0.07269,   Adjusted R-squared:  0.05935
F-statistic: 5.448 on 2 and 139 DF,  p-value: 0.005274
```





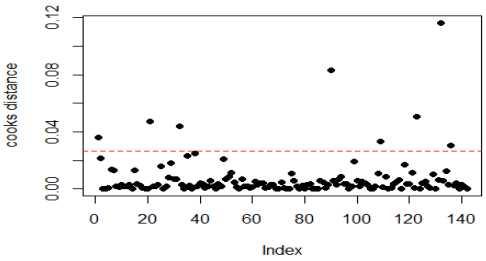
잔차분석한 결과, 변환전과 후가 크게 차이가 없는 듯하다. 또한 해석적인 측면에서도 변환전이 더 장점이 많으므로, 변환하지 않는 것이 더 좋다고 판단된다. 그러므로 변환하지 않는 모형을 최종모형으로 간주한다.

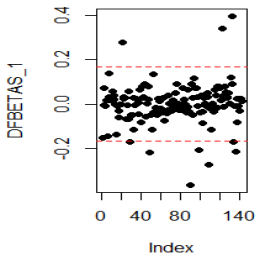
3. 4. 회귀진단· 모형확인 – 다중공선성

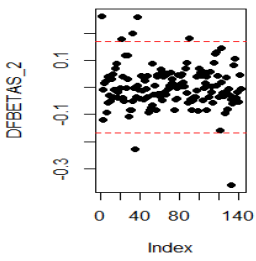
```
> vif(model2)
      x1       x2
1.008172 1.008172
```

5 보다 다 작으므로 다중공선성을 의심하지 않는다.

3. 5. 회귀진단· 모형확인 – 영향력 측도





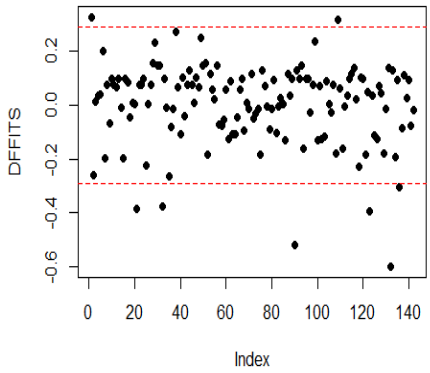


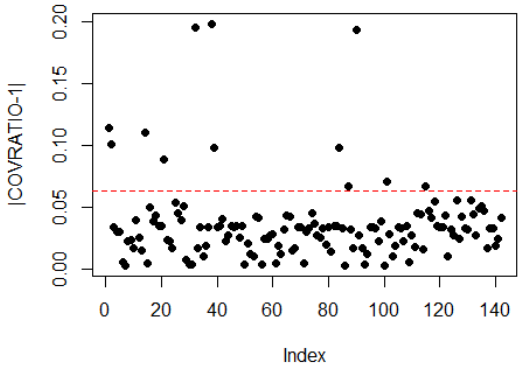
(1.) Cook's distance **(2.) DFBETAS**

(1) 이상치(8 개): 1, 21, 38, 90, 109,123, 132, 136

(2) BETA1 이상치(8 개): 21, 38,49, 90, 99, 109, 132, 136

BETA2 이상치(7 개): 1, 21, 32, 35, 38, 90, 132





(3.) DFFITS **(4.) COVRATIO**

(3.) 이상치(8 개): 1, 21, 32, 90, 109,123, 132, 136

(4.) 이상치(12개): 1, 2, 17, 21, 32, 35, 81, 86, 90, 93, 100, 118

총 4 가지 영향력 측도를 계산한 결과 공통적으로 추출된 이상치 8 개를 데이터에서 제외시켰다. 표본의 수는 142 명에서 134 명으로 줄어들었고, 더빈-왓슨 검정을 통해 통계량이 2 랑 비슷하므로 오차항의 독립성이 만족된다고 볼 수 있다.

```
> dwtest(Final_model)

        Durbin-Watson test

data:  Final_model
DW = 2.4102, p-value = 0.9929
alternative hypothesis: true autocorrelation is greater than 0
```

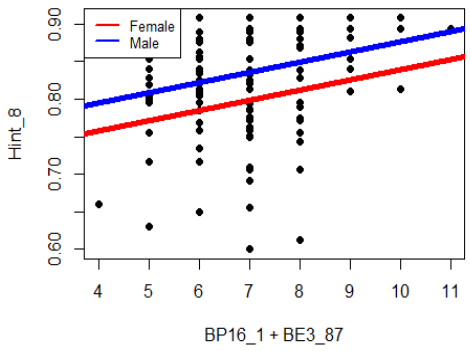
4. 모형확인

Train_data 를 70%, test_data 를 30%로 잡았고, PREE, R^2pedict SSE, R^2 값을 도출하였다.

```
> PRESS_Final$stat #PRESS
[1] 0.5271493
> SST <- sum((health_data$HINT_8-mean(health_data$HINT_8))^2)
> SSE <- sum(resid(Final_model)^2) #SSE
> SSR <- SST-SSE
> SSE
[1] 0.5072338
> 1-(SSE/SST) #R2
[1] 0.07910797
> 1-(PRESS_Final$stat/SST) #R2_predic
[1] 0.042951
```

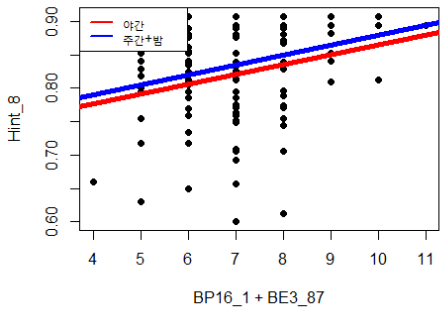
이상치를 제거한 모형은 PRESS 0.52 이고 SSE는 0.5이다. 서로 비슷하므로 새로운 자료에 대한 예측력의 정확도가 좋다고 평가할 수 있다. 그리고 R2_predict은 0.07이고 R2SMS 0.49이므로 이것 또한 서로 비슷하므로 적합도 측면에서는 좋지 않다고 평가할 수 있다. 즉, 이상치를 제거하기 전 세운 모형보다 예측모형과 적합도면에서 더 좋다고 평가할 수 있다.

5, 1. 성별과 HINT_8지수의 차이



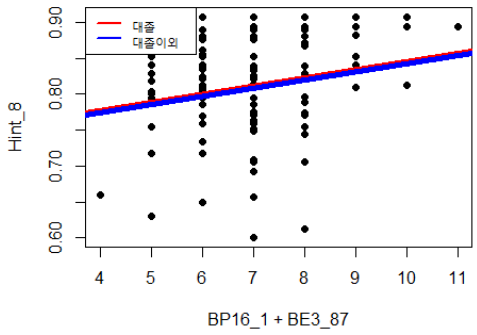

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.702300   0.036388  19.300  < 2e-16 ***
BP16_1       0.013702   0.005006   2.737  0.00706 **
BE3_87       0.010141   0.008059   1.258  0.21055
factor(sex)1 0.037125   0.021058   1.763  0.08023 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0691 on 131 degrees of freedom
Multiple R-squared:  0.0928,    Adjusted R-squared:  0.07202
F-statistic: 4.467 on 3 and 131 DF,  p-value: 0.005075
```

5.2 근무형태과 HINT_8지수의 차이



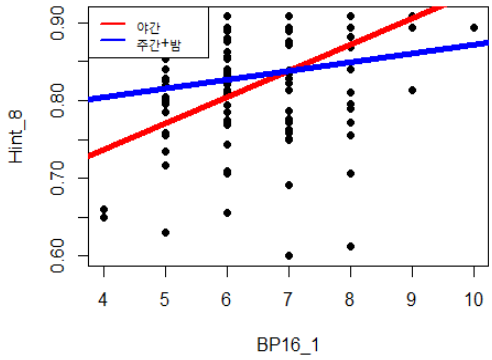

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.721840   0.051408  14.041  < 2e-16 ***
BP16_1            0.014762   0.005027   2.936  0.00392 **
BE3_87            0.010230   0.008201   1.247  0.21450
factor(EC_wht_5)1 0.007551   0.041069   0.184  0.85441
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06991 on 131 degrees of freedom
Multiple R-squared:  0.07151,   Adjusted R-squared:  0.05025
F-statistic: 3.363 on 3 and 131 DF,  p-value: 0.02075
```

5.3 교육수준과 HINT_8 지수의 차이




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.730600   0.034125  21.410  < 2e-16 ***
BP16_1       0.014730   0.005031   2.928  0.00402 **
BE3_87       0.010700   0.008252   1.297  0.19702
factor(edu)1 -0.002898  0.012196  -0.238  0.81253
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0699 on 131 degrees of freedom
Multiple R-squared:  0.07167,   Adjusted R-squared:  0.05042
F-statistic: 3.371 on 3 and 131 DF,  p-value: 0.02053
```

5.4 HINT8지수에 대한 주중수면시간과 근무형태의 교호작용




```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.94064    0.20920   4.496 1.51e-05 ***
BP16_1                   -0.01979    0.03241  -0.610    0.543
factor(EC_wht_5)1        -0.20986    0.21187  -0.990    0.324
BP16_1:factor(EC_wht_5)1  0.03519    0.03281   1.073    0.285
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07001 on 131 degrees of freedom
Multiple R-squared:  0.06867,   Adjusted R-squared:  0.04734
F-statistic:  3.22 on 3 and 131 DF,  p-value: 0.02493
```

5.5 해석

성별과 HINT_8 지수는 남녀별로 약간의 차이가 보여지고, 근무형태 또한 야간근무자와 주간+ 저녁 근무자 HINT_8 지수와 아주 약간의 차이가 보여지고, 교육수준은 차이가 나지 않는 것으로 보인다. 또한 HINT_8 지수에 대한 주중수면시간과 근무형태의 교호작용은 7 를 기준으로 교차를 한다. 하지만 factor(EC_wht_5)의 t-value 는 1.073 이므로 엄청나게 효과가 있는 것은 아니고 미미하다고 판단되어진다.

III 결론

factor(EC_wht_5)의 t-value는 1.073이므로 주중수면시간 및 근무형태가 HINT_8지수에 엄청난 영향을 주는 것은 아니다. 필자는 서론에서 수면관리의 중요성을 강조하기 위해서 수면관리와 근무형태가 유의미할 것이라고 생각을 하고 데이터분석을 하였다. 하지만 설명변수를 너무 적게 설정하여 R^2값의 유의미한 결과를 도출할 수 없었고 변수를 선택할 때 p-value 값을 기준으로 많이 한 것 같아서 변수가 유의미하지 않을 수도 있을 것이다. 또한 필자가 실수한 부분이 미리 범주형 설명변수를 넣고 회귀분석을 했어야 했는데 회귀분석 강의안을 따라서 하느라 이 부분이 나중에 넣어야 되는 줄 알고 착각한 부분이 너무 아쉬웠다.

현재 삼성서울병원·성균관대·카이스트 연구팀에서 교대 근무 간호사들을 대상으로 수면 패선 분석을 한 결과, “야간 근무 후 짧게, 주만 근무 후 길게 수면을 취하면 주간 졸림이 완화”된다는 결과를 도출하였다. 현재 이렇듯 많은 연구팀에서 수리모델을 이용해 개인 맞춤형 수면 패턴을 실시간으로 제공하고 있다. 필자도 열심히 데이터분석 실력을 향상시켜서 정신적으로 고통받고 있는 그들에게 힘을 주고 싶다. 또한 데이터분석을 혼자서 하는 것은 처음이었는데 나의 실력이 많이 부족하다는 것을 뼈빠지게 알게 되었다. 이것을 발판 삼아 데이터 분석에 관심을 갖고 공부해서 멋진 데이터분석가로 거듭나는 사람이 되고 싶다.

“

Ⅳ 참고문헌

질병관리본부, 한국형 건강관련 삶의 질 측정도구(HINT-8)의 가치평가 연구 , 주관연구기관: 울산대학교 산합협력단, 2017.04.19

정규직여부가 건강행태, 정신건강 및 삶의 질에 미치는 영향 -제6기(2013년) 국민건강영양조사 자료활용, 김성은 등 10명, 대한스트레스학회, 2016

세계일보, “교대근무자, 야근 후 짧게 주근 후 길게 자면 ‘주간졸림’ 완화”, 2021.10.06

오마이뉴스, “당신의 밤은 건강하십니까? 야간작업과 건강 관리”, 2021.12.01