회귀분석 보고서

수면의 질에 영향을 미치는 요인분석

10조 : 서재완, 서유찬, 이도권, 옥명주, 윤주영

개요

Ⅰ. 서론

Ⅱ. 본론

1) 데이터 전처리

2) 자료분석

(1) 다중회귀모형 적합

(2) 다중회귀모형 수정

(3) 단순회귀분석(걸음 수와 수면의 질)

Ⅲ. 결론

Ⅰ. 서론

1. 분석 동기

수면은 신체적, 정신적 피로를 회복하고 집중력과 기억력을 향상시키는 등의 역할을 담당한다. 동물의 신체는 에너지를 지속적으로 소비하면 그 기능이 떨어지며 피로를 느끼는데, 잠을 자는 동안에 일부 기관을 제외한 신체가 일을 쉬면서 떨어진 기능을 회복하고 피로를 회복한다. 잠이 우리 인생의 약 3분의 1을 차지하는 만큼 수면의 질은 대단히 중요하다고 말할 수 있지만 하지만 현대인의 과반수가 스트레스 등 다양한 이유로 수면 장애를 겪고 있다는 조사 결과가 있다. 최종 목적은 수면의 질을 올릴 수 있는 요인을 찾고 개인에게 맞는 진단을 하는 것이다.

2. 데이터 수집방법

sleep cycle이라는 수면 모니터링 앱을 통하여 얻었다. 이 앱은 사용자의 수면 패턴을 추적한다. 사운드 및 진동 분석을 통하여 심박수 등의 몸상태에 따른 데이터와 커피, 스트레스 등의 수면의 질에 미치는 효과를 기록하고 수면 데이터를 수집하여 수면에 대한 통계를 내리거나 수면 그래프를 그린다.

3. 데이터 분석 과정

수면의 질을 향상시킬 수 있는 방법을 찾기 위해서 수면의 질을 종속변수로 설정하고, 수면의 질에 영향을 미칠 것이라고 생각한 여러가지 요인들을 독립변수로 설정한다.

start : 잠에 든 시간, **end :** 잠에서 깨어난 시간, **sleep quality :** 수면의 질,

time in bed : 총 수면 시간, **wake up :** 잠에서 깨어난 후 느낀 기분,

sleep notes : 커피, 스트레스, 운동, 야식 등이 수면의 품질에 미치는 효과를 기록한 것,

heart rate : 심박수, **activity :** 걸음 수.

5년동안 측정 앱을 사용하여 얻어진 한 사람의 데이터를 통하여 이 사람에게 있어 수면에 영향을 미치는 요인은 무엇인지 분석한다. 그리고 분석된 데이터를 통해서 수면의 질을 향상시킬 수 있는 요인을 찾고 개인화된 진단을 하는 것이 이번 프로젝트의 최종 목표이다.

Ⅱ. 본론

- **Raw data 전처리**

1. Start와 End

Raw data에서는 YYYY-MM-DD HH:MM:SS 와 같은 형식으로 Start time과 End time이 표기되어 있다. 분석을 진행할 때 날짜는 고려 대상이 아니었기 때문에 제거했다. 좀 더 세밀한 분석을 위해서, 시각을 기준으로 하지 않고, 분을 기준으로 하여 진행했다. 분을 기준으로 변환 하였을 때, 자정을 넘기는 순간 1439분(23시 59분)에서 0분으로 값이 떨어지는 문제점이 있었다. 이를 위해서 가장 일찍 잠에 든 시각인 20:14을 기준으로 하여 Start와 End의 time을 변환하였다.

2. Sleep quality

Sleep quality의 %를 제거하여 0~100사이의 숫자로 나타냈다.

3. Time in bed

Start와 End의 전처리와 마찬가지로 세밀한 분석을 위해 HH:MM과 같은 형식을 따르고 있는 Time in bed 변수를 분단위로 나타냈다

4. Wake up

Wake up 변수는 :) , :| , :(와 같은 3가지 항목으로 이루어져 있다.

Wake up이 기록되지 않은 부분은 평상시로 생각하여 (wp_well:0, wp_bad:0), :)인 경우는 wp_well:1, :| 와 :(한 경우는 묶어서 wp_bad:1로 더미변수화 하였다. **:(**는 전체 데이터 중 하나밖에 없어서 이런 방법으로 처리했다.

5. Sleep Notes

Sleep Notes의 항목은 Stressful day, Drank coffee, Drank tea, Worked out으로 총 4가지 항목으로 이루어져 있다. 4가지 항목들을 각각 column으로 나타내고, 이 항목들이 각 행의 Sleep Notes 데이터에 해당하는 경우에 1로 표기한다. 해당하지 않는 경우에는 0으로 표기한다. 각 항목을 별개의 변수로 다룸으로써 수면의 질에 영향을 미치는 요인의 후보군을 다양화하였다.

- **결측치 처리 방법**

총 데이터 수는 887행이고, 그 중 결측치 수는 다음과 같다.

Start	End	Time.in.bed	Heart.rate	Wake.up	Activity (steps)	Sleep notes
0	0	0	725	641	418	209

많은 결측치는 회귀분석을 진행하는데 있어 걸림돌이 되기에 아래와 같은 방법을 이용하여 결측치를 처리했다.

1) Heart rate, Wake up 데이터에서 중간중간 데이터의 결측치가 존재하며, 이를 평균으로 대체했다.

Heart rate는 244번 데이터, Wake up은 250번 데이터 이후로는 값이 존재하지 않으므로 0~250행까지의 데이터만 따로 분리하여 다중회귀분석을 진행했다.

2) Activity (steps)

413번 데이터 이전에는 결측치를 가지고 있고, 그 이후의 데이터부터 값이 존재했다. 따라서 413행 이후의 데이터만 따로 분리하여 단순 회귀로 분석을 진행했다.

3) Sleep Note

678행 이후의 데이터가 모두 결측치이므로, 해당하는 행들을 모두 제거했다.

결측치 처리를 통해 얻은 2가지 파일 중 Heart rate와 Wake up의 데이터가 존재하는 0~250행까지의 데이터 파일로는 다중회귀분석을 진행하였고, Activity (steps)의 데이터가 존재하는 413~678행의 데이터 파일로는 단순회귀분석을 진행하였다.

- **전처리한 데이터로 회귀모형 적합**

Sleep.quality ~ All variable

회귀식의 정보를 요약해본 결과 Time. In. bed가 유의한 변수인 것을 확인할 수 있었고, Adjusted R-squared로 모형이 0.614의 설명력을 가지며 p-value 확인했을 때 회귀모형의 유의함을 확인하였다.

R^2adj	P_value	선형성	정규성	등분산성	독립성
0.614	< 2.2e-16	**X**	**X**	**X**	**O**

하지만 회귀가정을 1가지만 만족하므로 모형의 수정이 필요하다.

- **회귀모형의 수정**

전처리가 된 raw 데이터로 회귀모형을 만들었을 때 독립성을 제외한 모든 가정이 맞지 않으므로 수정이 필요하다. **다중 공선성 확인 > 변수선택 > 변수 변환 > 이상치, 영향관측치 처리** 순으로 진행하였다.

1) 다중 공선성 확인

다중 공선성을 확인했을 때, wp_well과 wp_bad가 8에 가까운 높은 값을 가진다. 따라서 변수 선택 시 둘 중 하나를 제거할 필요가 있다.

Start	End	Time.in.bed	Heart.rate	Stress	coffee	tea	Work.out	Ate.late	Wp.well	Wp.bad
1.8	1.2	1.7	1	1		1	1	1	7.8	7.8

2) 변수선택

모형의 해석을 적절하게 하기 위해 설명력이 떨어지더라도 독립 변수를 줄이는 것이 낫다고 판단하였다. **Backwards elimination**과 **stepwise method**을 사용하여 변수 선택을 진행하였고 동일한 결과를 얻었다.

Sleep.quality ~ Start +Time.in.bed + Drank.coffee + Drank.tea + wp_well

위처럼 독립변수를 구성했을 때 AIC가 1100.22로 가장 작았다. 그리고 wp_bad가 제거됨으로써 다중 공선성문제를 해결되었다. 따라서 회귀모형은 이 변수들로 구성한다.

하지만 여전히 회귀 가정을 만족하지 않아서 변수 변환도 진행하였다.

R^2adj	P_value	선형성	정규성	등분산성	독립성
0.6225	< 2.2e-16	X	X	X	O

3) 변수 변환

데이터 특성을 파악해보았을 때 변수들이 상당히 **skewed** 한 상태임을 알 수 있다. 이를 해결하기 위해서 종속변수를 변환하기로 하였다. 일반적으로는 주로 log변환을 사용하지만, 종속변수의 skew가 음수이므로 log변환을 하면 식이 복잡해져 해석이 어려워진다는 문제점이 있다. 프로젝트의 주목적은 예측이 아닌 설명이므로 해석이 어려운 모형을 사용하는 것은 적절하지 않다. 그래서 **powertransform**부록10을 사용하여 변수를 변환하였다. 종속변수를 **sq_transform = sleepquality^2.3712**으로 재정의하니 -2.16였던 왜도 값이 -0.15로 줄어들어 정규분포를 따르게 되고 skew 정도가 어느정도 해소되었다.

Summary를 보면 설명력이 0.6225에서 0.4301로 다소 감소했으나 통계적으로 여전히 유의하며 정규성과 독립성은 만족되는 모습을 보인다.

R^2adj	P_value	선형성	정규성	등분산성	독립성
0.4301	< 2.2e-16	X	O	X	O

나머지 가정들도 만족시키기 위해 이상치, 영향관측치를 처리하고자 한다.

4) 이상치, 영향관측치 처리

- **Outlier candidate : 53**

	Sleep.quality	Start	Time.in.bed	Drank.coffee	Drank.tea	Wp_well
53	54	39	559	1	1	1

53 case는 잠을 잔 시간은 평균(450분)보다 많고 나머지 독립변수들도 drank.coffee를 제외하고는 일반적으로 수면의 질이 좋을 때의 조건과 비슷하다. 하지만 수면의 질은 평균(76)보다 한참 낮은 54여서 이상치 후보이다. 하지만 Bonferroni p > 0.05 여서 무작정 제거하는 것은 좋지 않다.

- **Influential : 2, 138**

	Sleep.quality	Start	Time.in.bed	Drank.coffee	Drank.tea	Wp_well
2	3	63	16	0	0	0
138	3	211	15	1	0	0

2, 138 case는 수면의 질이 3으로 평균(76)보다 한참 작아서 영향관측치가 되었다. 잠을 잔 시간이 16,15로 상당히 작아서 수면의 질이 낮은 이유가 타당하므로 함부로 제거할 수 없다.

- **이상치, 영향관측치 제거 상황에 따른 회귀모형 특성 변화**

remove	R^2adj	P_value	선형성	정규성	등분산성	독립성
53	0.4384	< 2.2e-16	X	O	O	O
2	0.4057	< 2.2e-16	X	O	X	O
138	0.408	< 2.2e-16	X	O	X	O
2 & 53	0.4269	< 2.2e-16	X	O	O	O
53 & 138	0.4294	< 2.2e-16	x	O	O	O

2 & 138	0.4011	< 2.2e-16	**O**	**O**	**X**	**O**
all	0.4246	< 2.2e-16	**O**	**O**	**X**	**O**

위 표를 보았을 때 53(이상치)을 제거했을 때 최대한 회귀가정을 만족시키면서 R^2adj 이 가장 높은 값을 가지므로, **이상치만 제거한 모형을 최종모형으로 설정**한다. 선형성을 만족시키기 위해 영향관측치를 제거하더라도, 등분산성이 깨지고 모형의 목적인 설명에 맞지 않으므로 영향관측치를 제거하지 않고 최종모형을 선택하기로 한다.

5) 최종 모형

sq_transform ~ Start + Time.in.bed + Drank.coffee + Drank.tea + wp_well , data = data1[-53,]

선형성	정규성	등분산성	독립성
X	**O**	**O**	**O**

- **Sleep Quaility와 Activity Steps 간의 선형 관계**

Sleep.quality ~ Activity Steps

선형성	정규성	등분산성	독립성
O	**X**	**X**	**고려X**

위의 회귀식이 3가지 가정을 다 만족하지 않아서 powerTransform 함수를 이용하여 Sleep Quality를 지수 변환하였다.

선형성	정규성	등분산성	독립성
O	**X**	**O**	**고려X**

sq_transform ~ Activity Steps

정규성은 만족하지 않으나 CLT에 의해 표본의 수가 증가했을 때 만족할 것이다. Sleep Quality ~ Activity Steps로 단순 선형 회귀식을 적합해보니 R^2이 0.007362로 낮은 수치를 보이지만 P-value가 0.03396으로 통계적으로 유의한 수치를 보인다. R^2이 낮게 나온 이유는 Sleep Quality라는 종속 변수와 Activity Steps라는 설명변수 하나로만 이루어진 회귀식을 만들었기 때문이라고 볼 수 있다. 즉, 이 회귀식을 통해서 Activity Steps라는 변수가 Sleep Quality에 관련은 있지만 큰 영향은 미치지 않는다는 것을 알 수 있다.

Ⅲ. 결론

1) 다중선형회귀 해석

모형의 p-value가 < 2.2e-16로 유의수준 0.05에서 회귀모형은 통계적으로 유의하다.

다중회귀분석 결과 최종선택한 모형은 아래이다.

$$\boldsymbol{sq_{transform}} = -\mathbf{15106.032} + \mathbf{14.851}(\boldsymbol{start}) + \mathbf{102.330}(\boldsymbol{time.in.bed}) - \mathbf{4563}(\boldsymbol{drank.coffee}) + \mathbf{115.851}(\boldsymbol{drank.tea}) + \mathbf{2151.558}(\boldsymbol{wp_{well}})$$

$$drank.coffee = \begin{cases} 1\ if\ drank\ coffee\ before\ going\ to\ bed \\ 0\ if\ not\ drank\ coffee\ before\ going\ to\ bed \end{cases}$$

$$drank.tea = \begin{cases} 1\ if\ drank\ tea\ before\ going\ to\ bed \\ 0\ if\ not\ drank\ tea\ before\ going\ to\ bed \end{cases}$$

$$wp.well = \begin{cases} 1\ if\ feel\ good\ after\ waking\ up \\ 0\ if\ feel\ the\ same\ as\ usual\ after\ waking\ up \end{cases}$$

$\boldsymbol{sq}transform = sleep.quality$^2.3712

회귀식에서 통계적으로 유의한 변수들만으로 회귀식을 설명한다. 유의한 변수들은 time.in.bed와 drank.coffee 두가지이다. Sq_transform은 수면의 질을 변환한 변수인데 우리가 궁금한 것은 sleep.quality와 다른 설명변수 사이의 관계이다. 따라서 X_j가 한단위 변하면 sleep.quality는 $\sqrt{|B_j|}$만큼의 비율이 변화한다고 해석할 것이다.

위 회귀식을 봤을 때 수면시간이 1분 증가할 때마다 수면의 질은 평균적으로 $102.33^{\frac{1}{2.3712}} = 7.0416$ 만큼 증가함을 보이고 있다. 또한 자기 전 커피를 마신 날은 커피를 마시지 않은 날보다 평균적으로 $4563^{\frac{1}{2.3712}} = 34.93$만큼 수면의 질이 감소한다.

$R^2adj = 0.4413$으로 5개의 독립변수들이 종속변수의 다름을 약 44.1% 설명할 수 있다. 이는 수면 앱이 측정할 수 있는 변수들의 종류에는 한계가 있어서 설명변수가 그리 높지 않은 것으로 보인다. 차후 앱이 개선되면 더 높아질 가능성이 있다.

2) 수면의 질과 전날 걸음수의 관계

P-value는 0.05보다 낮지만 R^2=0.02654로 낮아서 해당 모형의 설명력이 너무 떨어진다. 따라서 걸음 수와 수면의 질 사이에 관계가 높지 않다.

위의 모형은 우리에게 몇 가지 부분을 시사한다. 수면 측정 앱으로 측정한 이 개인의 수면의 질은 수면시간과는 양의 상관관계를 가지며 커피는 음의 상관관계를 가진다. 결과적으로 수면시간을 늘리고 커피를 되도록이면 줄이는 것이 이 사람이 수면의 질을 높이는데 도움이 될 수 있음을 알 수 있다. 한 사람에 대한 회귀분석이라 큰 의미가 있을까라는 의문이 들 수 있으나 수면의 질에 영향을 미치는 요인은 개인에 따라 다르며 각각의 개인에 따라 다른 해결책을 줄 수 있다는 점에서 유용하다고 본다. 또한 이 수면 앱에서 다른 주요요인(수면장소, 심리상태)까지 측정이 가능하다면 회귀모형의 설명력이 더 좋아질 수 있다. 이는 앱의 또 다른 개선방향도 생각해볼 수 있다. 이는 서론에서 얘기한 개인화된 진단이라는 목적에도 부합한다.

<부록>

[그림 1] Raw data

	Start	End	Sleep quality	Time in bed	Wake up	Sleep Notes	Heart rate	Activity (steps)
0	2014-12-29 22:57:49	2014-12-30 07:30:13	100%	8:32	:)		59	0
1	2014-12-30 21:17:50	2014-12-30 21:33:54	3%	0:16	:|	Stressful day	72	0
2	2014-12-30 22:42:49	2014-12-31 07:13:31	98%	8:30	:|		57	0
3	2014-12-31 22:31:01	2015-01-01 06:03:01	65%	7:32				0
4	2015-01-01 22:12:10	2015-01-02 04:56:35	72%	6:44	:)	Drank coffee:Drank tea	68	0

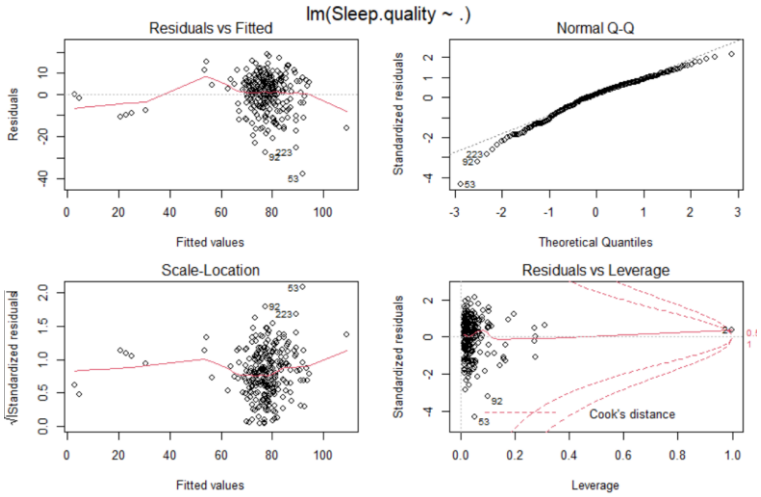
[그림 2] Heart rate와 Wake up의 데이터가 존재하는 0~250행까지의 데이터(전처리 후)

	Start	End	Sleep.quality	Time.in.bed	Heart.rate	Activity..steps.	Stressful.day	Drank.coffee	Drank.tea	Worked.out	Ate.late	wp_well	wp_bad
0	163	600	100	512	59.0	0	0	0	0	0	0	1	0
1	63	1443	3	16	72.0	0	1	0	0	0	0	0	1
2	148	583	98	510	57.0	0	0	0	0	0	0	0	1
3	137	513	65	452	60.6	0	0	0	0	0	0	0	0
4	118	446	72	404	68.0	0	0	1	1	0	0	1	0

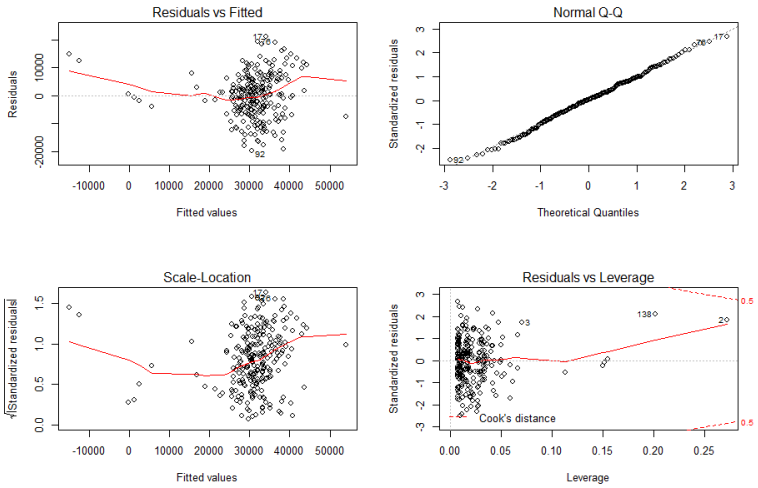
[그림 3] Activity (steps)의 데이터가 존재하는 413~678행의 데이터(전처리 후)

	Start	End	Sleep.quality	Time.in.bed	Wake.up	Heart.rate	Activity..steps.	Stressful.day	Drank.coffee	Drank.tea	Worked.out	Ate.late	wp_well	wp_bad
0	148	525	78	453	NaN	NaN	7200	0	1	0	1	1	0	0
1	104	533	36	504	NaN	NaN	3444	0	1	0	1	0	0	0
2	110	663	56	628	NaN	NaN	7901	0	1	0	0	0	0	0
3	114	530	52	491	NaN	NaN	3786	0	1	1	0	0	0	0
4	119	532	35	489	NaN	NaN	2668	0	1	0	1	0	0	0

[그림 4] sleep.quality ~ all variable 회귀 가정 검증



[그림 5] 최종 모형의 회귀 가정 검증



[그림 6] 최종 모형의 summary

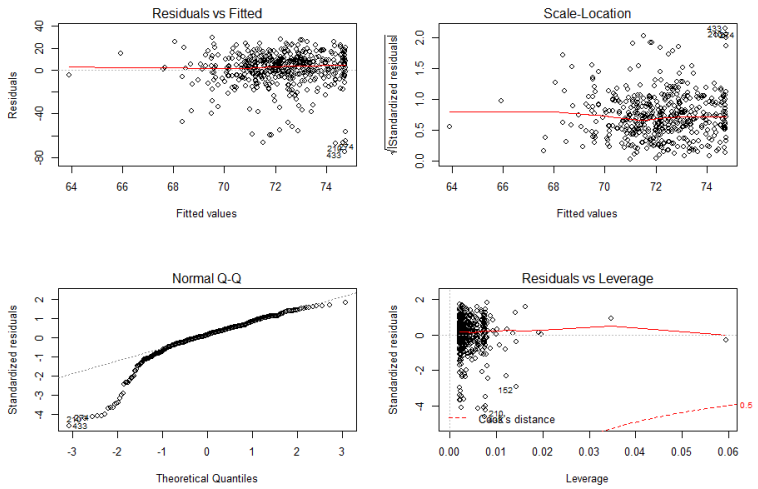
```
Call:
lm(formula = sq_transform ~ Start + Time.in.bed + Drank.coffee +
    Drank.tea + wp_well, data = data1[-53, ])

Residuals:
     Min       1Q   Median       3Q      Max
-22558.3  -5634.6    267.3   6483.4  24378.2

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17479.481   5295.028  -3.301 0.001108 **
Start            17.132      9.391   1.824 0.069326 .
Time.in.bed     116.946      9.660  12.106  < 2e-16 ***
Drank.coffee  -5270.457   1574.564  -3.347 0.000946 ***
Drank.tea       107.791   1290.715   0.084 0.933513
wp_well        2467.903   1702.734   1.449 0.148522
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9144 on 243 degrees of freedom
Multiple R-squared:  0.4497,    Adjusted R-squared:  0.4384
F-statistic: 39.72 on 5 and 243 DF,  p-value: < 2.2e-16
```

[그림 7] Activity steps 회귀검증



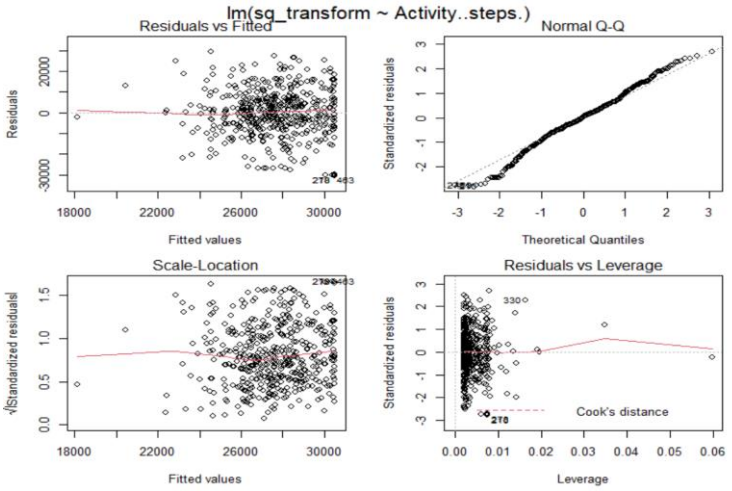
[그림 8] Activity steps ~ sleepquality 최종모형

```
Call:
lm(formula = sq_transform ~ Activity..steps., data = data2[-c(433),
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-30444.0  -6233.4    187.8   6667.9  29405.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      30514.0349   967.9418  31.525  < 2e-16 ***
Activity..steps.    -0.5663     0.1590  -3.562 0.000405 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11060 on 473 degrees of freedom
Multiple R-squared:  0.02613,   Adjusted R-squared:  0.02407
F-statistic: 12.69 on 1 and 473 DF,  p-value: 0.000405
```



[10] Powertransform

검정력 변환(powerTransform)은 Box-Cox(1964)의 최대우도 유사 접근 방식(maximum likelihood-like approach)을 사용하여 정규성, 선형성 및/또는 상수 분산에 대한 일변량 또는 다변량 반응의 변환을 선택한다.

최대우도추정(MLE)은 통계학에서 일부 관측된 데이터가 주어졌을 때 가정된 확률분포의 매개변수를 추정하는 방법이다.

참고자료

1) raw데이터(personal sleep data from sleep cycle ios app)

   https://www.kaggle.com/danagerous/sleep-data

2) powertransform

   powerTransform function - RDocumentation

3) sleepcycle 앱 사이트

   How Sleep Cycle Works: Sleep Tracker & Alarm Clock User Guide