moderation

kang hyeon gu

2022 9 1

사용한 데이터는 각 사람들의 나이(age), 성별(sex), bmi지수, 자식 수(children), 흡연 여부(smoker), 사는 지역(region)에 따른 보험료(charges)를 나타내고 있습니다. 이 데이터를 이용해 bmi지수에 따라 보험료의 비용이 어떻게 달라지는지 알아보기 위해서 데이터 분석을 실시하였습니다. 입력변수(input variable)로 bmi지수 출력변수(outcome variable)는 보험료로 설정했습니다. 그리고 이때 성별과 흡연 여부가 출력 결과에 영향을 미치는지 알아보았습니다.

data=read.csv('C:/R/rdata/insurance.csv')  
library(lattice)  
library(ggplot2)

library(dplyr)

# 데이터 분석을 위해서 성별이 남자면 0, 여자면 1, 흡연을 안 하면 0, 하면 1로 처리했습니다.

data$sex <- ifelse(data$sex=='male',0,1) # 남자 0 여자 1  
data$smoker <- ifelse(data$smoker=='yes',1,0) # 피면 1 안피면 0  
data$charges=data$charges/1000  
head(data)

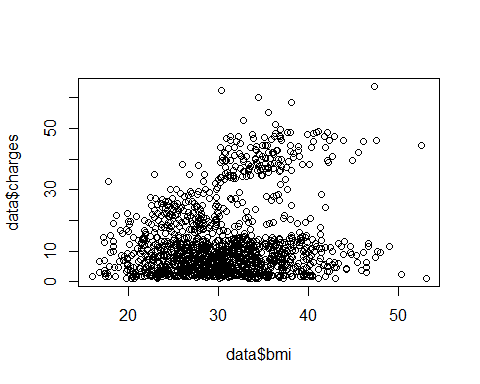
## age sex bmi children smoker region charges  
## 1 19 1 27.900 0 1 southwest 16.884924  
## 2 18 0 33.770 1 0 southeast 1.725552  
## 3 28 0 33.000 3 0 southeast 4.449462  
## 4 33 0 22.705 0 0 northwest 21.984471  
## 5 32 0 28.880 0 0 northwest 3.866855  
## 6 31 1 25.740 0 0 southeast 3.756622

cor(data[1:5],data$charges)

## [,1]  
## age 0.29900819  
## sex -0.05729206  
## bmi 0.19834097  
## children 0.06799823  
## smoker 0.78725143

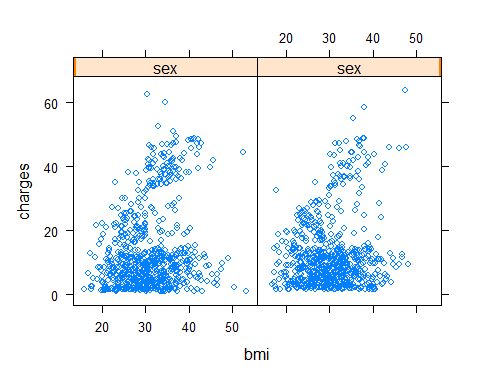
plot(data$bmi,data$charges)

데이터 분석을 하기에 앞서 bmi지수에 따른 보험료와 여기에 moderatorvariable로 성별, 흡연 여부에 따라 보험료가 어떻게 달라지는지 산점도를 그려보았습니다.



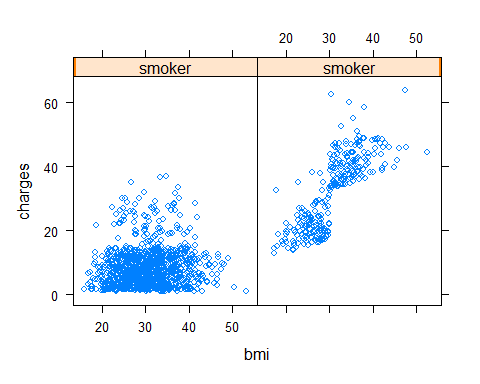
Bmi 지수가 커질수록 보험료가 커지는 양상을 볼 수 있습니다. 회귀분석 결과는 4페이지에 있습니다.

xyplot(charges~ bmi | sex, data= data )



성별에 따른 산점도를 살펴본 결과 bmi지수에 따른 보험료에 크게 영향을 미치지 않는 것 같습니다. 분석 결과는 7페이지에 있습니다.

xyplot(charges~ bmi | smoker, data= data )



흡연여부가 bmi지수에 따른 보험료에 크게 영향을 미치는 것 같습니다. 분석 결과는 6페이지에 있습니다.

먼저 bmi지수에 따른 보험료가 어떻게 달라지는지 회귀분석을 해보았습니다.

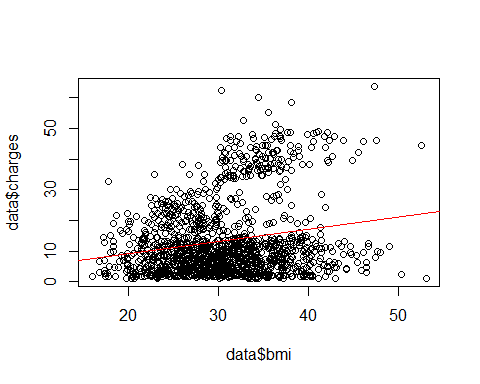
# bmi  
bm=summary(lm(formula=data$charges~data$bmi))  
bm$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.192937 1.66480238 0.7165639 4.737685e-01  
## data$bmi 0.393873 0.05325074 7.3965741 2.459086e-13

plot(data$bmi,data$charges)  
abline(1.192937,0.3993873,col='red')

이 회귀분석 모델은 다음과 같이 설명될 수 있습니다. Y=β0+β1∗X+ϵ

여기서 Y는 보험료, X는 bmi지수입니다. 회귀분석 결과 β1는 α 수준 0.05에서 p-value=2.459086e-1이므로 추정된 회귀계수는 유의합니다. 즉 β1은 0.393873이라 볼 수 있습니다. R-squared: 0.03934 이므로 회귀직선의 설명력이 높은 편은 아닙니다.



moderator variable로 흡연여부가 사용된다면 어떻게 달라지는 보겠습니다.

Y는 보험료, X는 bmi지수, Z는 흡연여부라고 하면 다음 모델을 사용하여 Moderation analysis를 할 수 있습니다.

Y=β0+β1∗X+β2∗Z+β3∗X∗Z+ϵ= β0+β1∗X+ϵFor non-smoker(Z=0)

β0+β2+(β1+β3)∗X+ϵFor smoker(Z=1)

# bmi smoker  
xz= data$bmi\*data$smoker  
smo=summary(lm(formula=data$charges~data$bmi+data$smoker+xz))  
smo

##   
## Call:  
## lm(formula = data$charges ~ data$bmi + data$smoker + xz)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.7680 -4.4007 -0.8695 2.9577 31.0559   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.87942 0.97687 6.019 2.27e-09 \*\*\*  
## data$bmi 0.08335 0.03127 2.666 0.00778 \*\*   
## data$smoker -19.06600 2.09203 -9.114 < 2e-16 \*\*\*  
## xz 1.38976 0.06678 20.810 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.161 on 1334 degrees of freedom  
## Multiple R-squared: 0.7418, Adjusted R-squared: 0.7412   
## F-statistic: 1277 on 3 and 1334 DF, p-value: < 2.2e-16

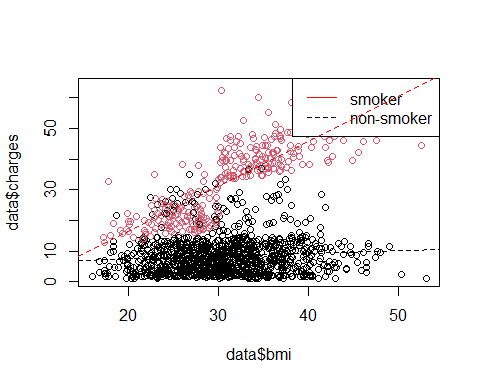
분석 결과 interaction term XZ에 대해 알파 수준 0.05하 p값이 p-value: < 2.2e-16

이므로 회귀계수 1.38976으로 추정할 수 있습니다. 즉 bmi지수에 따른 보험료는 흡연여부에 따라 크게 다릅니다. 흡연여부에 따른 회귀직선은 다음과 같습니다.

Y= 0.08335X + 5.87942 For non-smoker (Z=0)

Y= 1.47311 – 13.1865 For smoker (Z=1) 그래프로 그려보면 다음과 같습니다.

plot(data$bmi,data$charges,col=factor(data$smoker))  
legend('topright',c('smoker','non-smoker'),lty=c(1,2),col=c('red','black'))  
abline(5.87942,0.08335,lty=2)  
abline(-13.18658,1.47311,lty=2,col='red')



moderator variable로 성별이 사용된다면 어떻게 달라지는 간략하게 보겠습니다.

Y는 보험료, X는 bmi지수, Z는 성별이라고 하면 다음 모델을 사용하여 Moderation analysis를 할 수 있습니다.

Y=β0+β1∗X+β2∗Z+β3∗X∗Z+ϵ= β0+β1∗X+ϵFor male(Z=0)

β0+β2+(β1+β3)∗X+ϵFor female(Z=1)

# bmi sex  
xz2= data$bmi\*data$sex  
se=summary(lm(data$charges~data$bmi+data$sex+xz2))  
se

##   
## Call:  
## lm(formula = data$charges ~ data$bmi + data$sex + xz2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.378 -8.115 -3.819 4.773 48.914   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.80555 2.34432 -0.344 0.7312   
## data$bmi 0.47708 0.07432 6.420 1.89e-10 \*\*\*  
## data$sex 4.34936 3.32811 1.307 0.1915   
## xz2 -0.17996 0.10649 -1.690 0.0913 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.86 on 1334 degrees of freedom  
## Multiple R-squared: 0.0437, Adjusted R-squared: 0.04155   
## F-statistic: 20.32 on 3 and 1334 DF, p-value: 7.029e-13

분석결과 interaction term XZ2에 대해서 알파 수준 0.05하 p값이 0.0913이므로 추정된 회귀계수는 무의미 합니다. 즉 bmi지수에 따른 보험료는 성별에 따라 크게 다르다고 볼 수 없습니다.

정리하자면 bmi지수가 커짐에 따라 보험료가 약간 인상되는 경향이 있습니다. 이때 성별과 흡연여부가 moderator variable로 사용되었을 때 성별이 크게 영향을 미치지 않았지만 흡연여부는 크게 영향을 미치는 것을 확인할 수 있었습니다. 즉 흡연여부가 moderator로 고려될 수 있습니다.