Moderation analysis: insurance data

Jose M Sallan

23/5/2018

## Medical charges data

We have 1,338 observations of medical charges of patients in the US. The available variables are:

* **age**: age.
* **sex**: gender.
* **bmi**: body mass index.
* **children**: number of children.
* **smoker**: consumes tobacco regulary (yes, no).
* **region**: region where the patient lives.
* **charges**: medical charges (dependent variable).

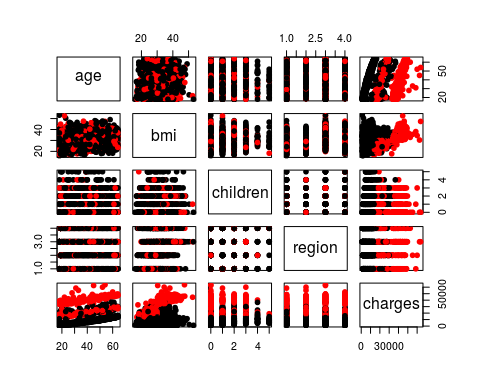
Let’s read the data:

insurance <- read.csv("insurance.csv")

## Examining data

Let’s examine relationships among variables plotting the dataset. We will obtain scatterplots for evey pair of variables. We will exclude gender and smoker, and we will plot in different colors smokers and non-smokers:

plot(insurance[ , c(1,3,4,6,7)], pch=19, col=insurance$smoker)

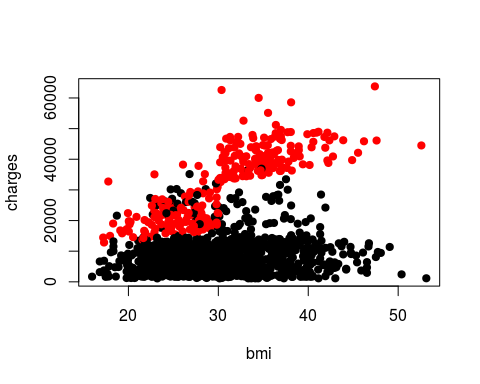


There seems to be an interesting relationship between charges and bmi dependent on smoker:

## Interaction effect

Let’s make a closeup of the bmi vs charges relationship:

plot(insurance$bmi, insurance$charges, pch=19, col=insurance$smoker, xlab="bmi", ylab="charges")



The plot reveals that it seems to be that bmi influences medical charges (i.e., your health) only if you smoke. Let’s examine the interaction effect:

bmi.interaction <- lm(charges ~ bmi\*smoker, data=insurance)  
summary(bmi.interaction)

##   
## Call:  
## lm(formula = charges ~ bmi \* smoker, data = insurance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19768.0 -4400.7 -869.5 2957.7 31055.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5879.42 976.87 6.019 2.27e-09 \*\*\*  
## bmi 83.35 31.27 2.666 0.00778 \*\*   
## smokeryes -19066.00 2092.03 -9.114 < 2e-16 \*\*\*  
## bmi:smokeryes 1389.76 66.78 20.810 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6161 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

In the resulting model, the interaction term is significant, confirming analytically the moderating relationship. As the moderating variable is binary, we can analyze the bmi and charges relationship for both values of the variable:

smokers <- insurance[which(insurance$smoker=="yes"), ]  
nonsmokers <- insurance[which(insurance$smoker=="no"), ]  
  
bmi.smokers <- lm(charges ~ bmi, data=smokers)  
bmi.nonsmokers <- lm(charges ~ bmi, data=nonsmokers)  
  
summary(bmi.smokers)

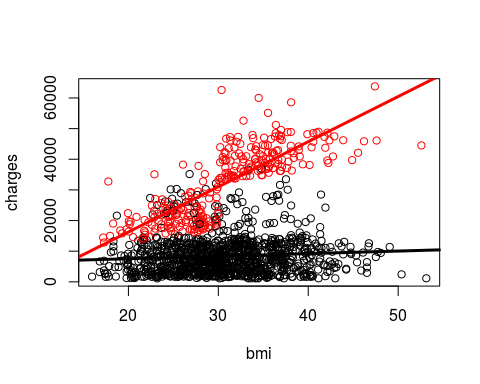
##   
## Call:  
## lm(formula = charges ~ bmi, data = smokers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19768.0 -4487.9 34.4 3263.9 31055.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -13186.58 2052.88 -6.423 5.93e-10 \*\*\*  
## bmi 1473.11 65.48 22.496 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6837 on 272 degrees of freedom  
## Multiple R-squared: 0.6504, Adjusted R-squared: 0.6491   
## F-statistic: 506.1 on 1 and 272 DF, p-value: < 2.2e-16

summary(bmi.nonsmokers)

##   
## Call:  
## lm(formula = charges ~ bmi, data = nonsmokers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9144 -4360 -1009 2922 28131   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5879.42 947.48 6.205 7.81e-10 \*\*\*  
## bmi 83.35 30.33 2.748 0.00609 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5975 on 1062 degrees of freedom  
## Multiple R-squared: 0.007062, Adjusted R-squared: 0.006127   
## F-statistic: 7.553 on 1 and 1062 DF, p-value: 0.006091

Finally the regression coefficients of both models are significant, but the one of the smoker’s model is larger than the one of the nonsmoker’s. Let’s plot both:

plot(insurance$bmi, insurance$charges, col=insurance$smoker, xlab="bmi", ylab="charges")  
abline(bmi.smokers, col="red", lwd=3)  
abline(bmi.nonsmokers, col="black", lwd=3)



## Interaction effect with the rest of variables

A common practice when reporting interaction effects is presenting two models: one model with the variables of interest, and the other model with the variables and the interaction term. Then, comparing the fit of both models we can check if the addition of the interaction adds explanatory power to the model. Let’s do it for that data:

model1 <- lm(charges ~ age + sex + bmi + smoker, data=insurance)  
model2 <- lm(charges ~ age + sex + bmi\*smoker, data=insurance)  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.1. https://CRAN.R-project.org/package=stargazer

stargazer(model1, model2, type="text")

##   
## =========================================================================  
## Dependent variable:   
## -----------------------------------------------------  
## charges   
## (1) (2)   
## -------------------------------------------------------------------------  
## age 259.453\*\*\* 266.372\*\*\*   
## (11.942) (9.612)   
##   
## sexmale -109.041 -473.495\*   
## (334.665) (269.612)   
##   
## bmi 323.051\*\*\* 7.969   
## (27.529) (25.044)   
##   
## smokeryes 23,833.870\*\*\* -20,193.150\*\*\*   
## (414.186) (1,666.491)   
##   
## bmi:smokeryes 1,435.608\*\*\*   
## (53.242)   
##   
## Constant -11,633.500\*\*\* -2,071.077\*\*   
## (947.267) (840.644)   
##   
## -------------------------------------------------------------------------  
## Observations 1,338 1,338   
## R2 0.747 0.837   
## Adjusted R2 0.747 0.836   
## Residual Std. Error 6,094.362 (df = 1333) 4,903.556 (df = 1332)   
## F Statistic 986.538\*\*\* (df = 4; 1333) 1,364.503\*\*\* (df = 5; 1332)  
## =========================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Data seem to confirm that the interaction effect is significant. Let’s assess it examining if model2 explains data better than model1:

anova(model1, model2)

## Analysis of Variance Table  
##   
## Model 1: charges ~ age + sex + bmi + smoker  
## Model 2: charges ~ age + sex + bmi \* smoker  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1333 4.9509e+10   
## 2 1332 3.2028e+10 1 1.7482e+10 727.04 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Data shows that we can discard the null hypothesis that both models are equivalent.