Linear Regression with Medical Insurance

Johann Piedras

11/28/2020

Contents

Which Factors Influce the Price Of Health Insurance?

First and foremost, this RMD file is inspired and founded thanks to a Python Notebook by mariapushkareva on Kaggle. The purpose here is to showcase an example of linear regression with medical data, but also to parallel data analysis from Python using R.

With that out of the way lets, look at the insurance.csv file sourced from Kaggle. Although I don't know exactly where Kaggle got the csv from, the purpose derived from the data is is to attempt to understand what factors contribute to the cost of health insurance.

EDA

```r
# importing the data
df <- read.csv('insurance.csv', header = TRUE)
head(df)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```