What are the effects of age, gender, numbers of children, being a smoker or not, living region on medical expenses?

Jingyi Cui

March 2020

Contents

1 Introduction

Background Information

As everyone knows, the healthcare is considerably expensive in United States. According to an annual survey, the cost of family health coverage in the U.S. now tops $20,000. Therefore, medical expenses is always a popular and attractive topic among people at all ages; there are a lot of news, reports, surveys about why the healthcare is such expensive in U.S. From my prospect, analyzing the medical cost should be divided into two parts: patients themselves and everything other than patients. We need to analyze this problem from those two parts; indeed, there is a host of information about the hospitals, doctors and nurses. In this project, I will start this problem with analyzing the patients themselves: whether some special features may have effects on their medical costs.

On one aspect, factors effecting the healthcare cost are wage of doctors and nurses, expensive mix of treatments, administrative costs, etc. On the other aspect, the patients' conditions may play an important role in the final medical costs: maybe older people's bill will be higher, since their disease may be complicated by other diseases and it requires much more time and much more expensive to do the treatments; maybe people who do sports everyday have a lower medical expense, since they may have a better condition and will recover soon, resulting low medical expense.

I once had an experience of using my insurance. In my second year of college life, my finger was injured badly and my final bill was around three thousand dollars, which is a huge amount. Fortunately, I was young and it only takes me several weeks to recover totally; besides, I do not drink or smoke. Doctors and nurses proved that not drinking alcohol or smoking will make me get well much more quickly My insurance covered most part of the bill: around two thousand dollars. Thus, I decide to use this topic for my ECON 483 project. [1]

Dataset

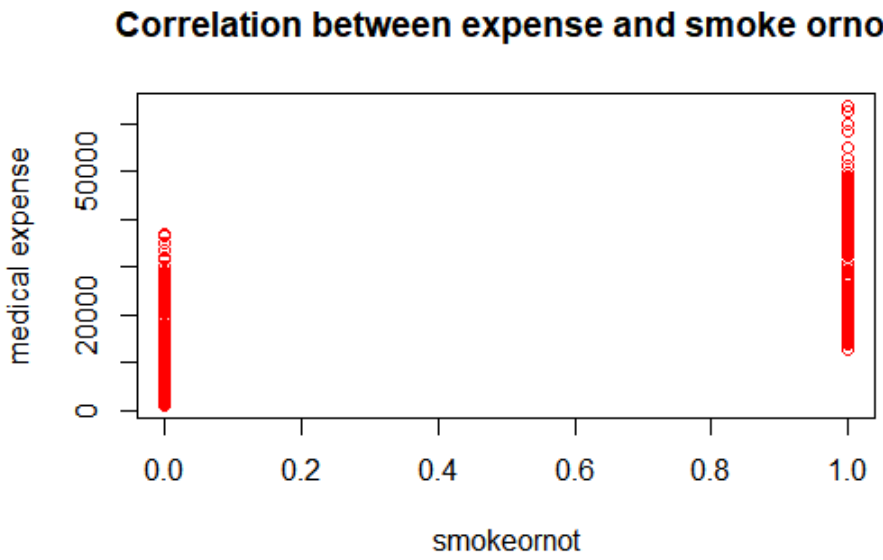
The dataset is from kaggle.com and it is clear dataset with 6 independent variables:

1. age: an integer number of the person.
2. gender: female or male.
3. bmi: Body Mass Index. A high BMI can be an indicator of high body fatness.
   $\text{BMI} = \frac{\text{weight(kg)}}{\text{height(m)}^2}$.
4. children: number of children in the insurance plan. It is an integer.
5. smoker: if smoking or not, a dummy variable.
6. region: A categorical variable and northeast, southeast, southwest and northwest in this dataset.

and 1 dependent variable:

1. charge: an numerical variables representing total spending in treatment.

[1] I changed my topic since while running the model for my last topic about subscribing the cable TV, there is no significance between variables



Just as we assume, smokers' medical expenses, around 10000 dollars to 60000 dollars, will be much higher than non-smokers, around thousands of dollars to 40000 dollars but most are below thirty thousand dollars, on average. Later, we will consider the effect of age and bmi on medical expenses for smokers and non-smokers, respectively.

The plots below are the relationship between age and expenses, bmi and expenses, for both non smokers and smokers. What we can observe from the plots is a positive trend between bmi and expenses for smokers, but not very clear relationship for non-smokers. In addition, the relationship between age and charges for both non-smokers and smokers is not clearly observable. Later, we will do regression considering those conditions.

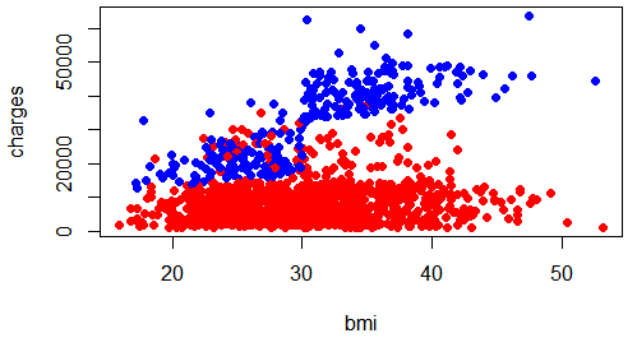


Figure (2) Blue: Smokers, Red: non-smokers

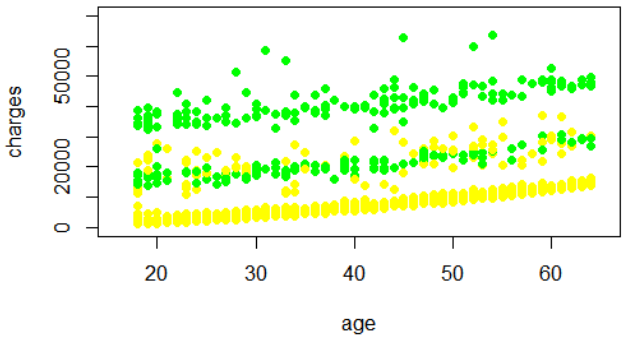


Figure (3) Green: smokers, Yellow: non-smokers

2 Model

1. charges $= \alpha + \beta_0 \cdot$ age $+ \beta_1 \cdot$sex $+ \beta_2 \cdot$bmi $+ \beta_3 \cdot$children $+ \beta_4 \cdot$smoker $+ \beta_4 \cdot$region $+ \epsilon$

```
> summary(fit)

Call:
lm(formula = charges ~ ., data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Figure 3: Summary of the model 1

We start with this basic model with all initial independent variables[2]: age, gender, bmi, number of kids in the insurance plan, being a smoker or not and living region. The independent variables of the linear model are statistically significant at 5% level and p-values are less than 0.05. What we get from the summary of the linear model is age, bmi, number of kids, being a smoker and living in southeast or southwest are significant. Holding all other variables fixed, if age increases by 1, the expense will increase by 256.9 dollars; if number of kids increase by 1, the final charges will increase by 475.5 dollars. In addition, adjusted R-squared is around 0.75, which means around 75% of variability can be explained by this model.

One thing I am surprised about this result is the coefficient in front the gender as male. Usually, people will think since males smoke more, they will have a worse health condition than female. However, right here, maybe because of lack of information and samples, this surprising result comes out.

2. charges $= \alpha + \beta_0 \cdot$ age $+ \beta_1 \cdot$sex $+ \beta_2 \cdot$bmi30[3]*smoker $+ \beta_3 \cdot$children $+ \beta_4 \cdot$smoker $+ \beta_4 \cdot$region $+ \epsilon$

   After doing the regression above, I consider the interaction effects of bmi30[4] and answer of being a smoker or not. The reason of using bmi30 rather than bmi is that 30 is a threshold value for the index and maybe only when one's bmi is above 30, bmi will have big effects on the medical expense; otherwise, there is no effect. This model will help us figure out among smokers and non-smokers, what will be the effect of bmi30 on the expenses.

[2] more variables will be added to the model later

[3] bmi30 is a specified index set by Centers for Disease Control and Prevention (CDC) and above 30.0 means obese in terms of weight status

[4] In this case, I set 1 for those whose bmi is larger than 30, others 0.

```
Call:
lm(formula = charges ~ age + children + bmi + sex + bmi30 * smoker +
    region, data = insurance)

Residuals:
    Min      1Q  Median      3Q     Max
-4081.3 -1830.2 -1263.2  -464.7 24813.0

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4622.33     953.77  -4.846 1.41e-06 ***
age               263.64       8.75  30.130  < 2e-16 ***
children          508.97     101.31   5.024 5.76e-07 ***
bmi               108.96      34.35   3.172 0.001549 **
sexmale          -470.14     244.98  -1.919 0.055185 .
bmi30            -803.06     423.19  -1.898 0.057964 .
smokeryes       13413.21     439.59  30.513  < 2e-16 ***
regionnorthwest  -263.74     350.20  -0.753 0.451514
regionsoutheast  -822.28     352.57  -2.332 0.019837 *
regionsouthwest -1165.72     351.45  -3.317 0.000935 ***
bmi30:smokeryes 19909.67     605.92  32.859  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4457 on 1327 degrees of freedom
Multiple R-squared:  0.8656,    Adjusted R-squared:  0.8646
F-statistic: 854.5 on 10 and 1327 DF,  p-value: < 2.2e-16
```

Figure 4: Summary of the model 2

3. charges $= \alpha + \beta \cdot$ age $+ \beta_1 \cdot age2$[5] $+ \beta_1 \cdot$sex $+ \beta_2 \cdot$bmi30*smoker $+ \beta_3 \cdot$children $+ \beta_4 \cdot$smoker $+ \beta_4 \cdot$region $+ \epsilon$

   Compared with last model, one independent variable is added. One thing we can consider is the relationship between the age and expense. We may assume it is non-linear: once one gets older, a larger amount of medical expenses may be charged compared with last year. Thus, we can add age^2 to do the further regression.

   Thinking about the summary of this model, we observe that a smoker whose bmi is larger than 30 spend much more than who only smoke or bmi above 30: one is both obese and smoking may need to spend 19912 dollars on average; one only smoke spends almost 14000 dollars for medical expenses per year on average. In addition, we can conclude that as people get older, they need to pay much more on medical costs.

[5] age2 is the squared values of age

```
lm(formula = charges ~ age + children + age2 + bmi + sex + bmi30 *
    smoker + region, data = insurance)

Residuals:
    Min      1Q  Median      3Q     Max
-4260.3 -1644.6 -1272.7  -784.7 24192.7

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        69.2494  1353.2349   0.051 0.959195
age               -21.6786    59.4956  -0.364 0.715638
children          661.5105   105.2784   6.283 4.48e-10 ***
age2                3.5978     0.7422   4.847 1.40e-06 ***
bmi               114.2920    34.0816   3.353 0.000821 ***
sexmale          -475.6760   242.9293  -1.958 0.050430 .
bmi30            -938.5116   420.5807  -2.231 0.025817 *
smokeryes       13421.6370   435.9158  30.790  < 2e-16 ***
regionnorthwest  -275.6659   347.2730  -0.794 0.427453
regionsoutheast  -826.1187   349.6181  -2.363 0.018275 *
regionsouthwest -1164.8152   348.5123  -3.342 0.000854 ***
bmi30:smokeryes 19912.6072   600.8493  33.141  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4419 on 1326 degrees of freedom
Multiple R-squared:  0.8679,    Adjusted R-squared:  0.8668
F-statistic: 792.1 on 11 and 1326 DF,  p-value: < 2.2e-16
```

Figure 5: Summary of the model 3

4. charges $= \alpha + \beta_0 \cdot$children $+ \beta_1 \cdot$region $+ \epsilon$

   In this time, I consider if omitting a few variables: smoke or not, gender, bmi and age, what happens to the living region and number of children in medical expenses in general?

```
lm(formula = charges ~ children + +region, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max
-13109  -8591  -4058   3107  49783

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      12660.8      728.5  17.379  < 2e-16 ***
children           712.5      273.8   2.603  0.00935 **
regionnorthwest  -1061.1      947.0  -1.120  0.26272
regionsoutheast   1326.8      920.9   1.441  0.14990
regionsouthwest  -1127.3      946.9  -1.190  0.23407
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12060 on 1333 degrees of freedom
Multiple R-squared:  0.01166,   Adjusted R-squared:  0.008691
F-statistic: 3.931 on 4 and 1333 DF,  p-value: 0.003539
```

Figure 6: Summary of the model 4

From the results the above, if smoke, gender bmi and age are omitted from the regression, living region gap and children gap tend t increase. And this is omitted variable bias. Particularly, smoke or not, gender, bmi and age are different across people living various regions and the number of kids in the insurance plan, which is the main source of omitted variable bias. According to the dataset, we can observe the people who have more kids in their insurance plan have larger bmi than individuals who have less kids covered in their insurance plan.

5. charges $= \alpha + \beta_0 \cdot bmi30 + \beta_1 \cdot smoke + \beta_2 \cdot bmi30 \cdot smoke + \epsilon$

   Finally, we consider using difference in difference to track the interaction between bmi30 and smoke or not.

```
lm(formula = charges ~ bmi30 + smoke + bmi30 * smoke, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max
 -9144  -4345  -1041   2968  28057

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      7966.9      262.2  30.386   <2e-16 ***
bmi30             886.3      361.1   2.455   0.0142 *
smokeyes        13402.3      578.6  23.165   <2e-16 ***
bmi30:smokeyes  19437.3      797.8  24.364   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5880 on 1334 degrees of freedom
Multiple R-squared:  0.7648,    Adjusted R-squared:  0.7642
F-statistic:  1446 on 3 and 1334 DF,  p-value: < 2.2e-16
```

Figure 7: Summary of the model 5

Observed from the result, all the variables are significant since their p values are all less than 0.05. For those who smoke and bmi is above 30, they will spend around twenty thousand dollars on medical expenses on average per year, holding all other variables constant. And this number is much higher than who just smoke and whose bmi is above 30 alone. This conclusion corresponds to what we get at first: for smokers, if bmi increases, medical costs will increase as well (in Figure (2)).

3 Results and Conclusion

1. Considering the basic model, age, bmi, number of kids in the insurance plan, being a smoker and living in southeast or southwest are significant variables. All these will have effects on the final medical expenses covered by the insurance plan. Except for being a male person and living in southwest, southeast, and northwest, all other variables' increment accompany with the increment of medical expenses on average.

2. As I separate the effect of bmi30 and bmi, the conclusion go further with more details. Bmi30 is significant variable affecting the medical expenses, meaning bmi below 30 will not have effects on the final charges. This leads to a question: among people whose bmi is above 30, if they smoke, what will happen to the result.

3. The interaction between bmi30 and smokeyes cannot be ignored. People who smoke and have large bmi usually have a worse health condition and may need to pay a much higher amount for medical costs, than people who only smoke. In addition, the relationship between age and charges is not linear: as people get older and older, they may need to pay much more than compared with a prior year, since they may have more health issues to be considered.

4 Limitations

1. Since the data is from kaggle.com, there is no enough data to do a better analysis. There is not enough information whether all these people have the same insurance plan of the same company, since the standard of covering is different across different companies and different insurance types in United States.

2. A part of results is different from what I thought. From example, the amount a male person need to pay less on average per year. And I guess the reason may be the small amount of samples. There are only 1300 observations, but people who own insurance and had used insurance before in United States are much more than 1300.

3. Due to the small data, we do not have enough information about policy for helping people whose age is above 65 or 70. And we do not how their costs are billed to both insurance firms and the government. More policies and relevant information are needed to be added to the data set.

4. In this dataset, there are only smokers. However, I believe people drinking or not may also have an effect, because people who drink alcohol may spend much more time to fully recover, resulting in higher medical costs at last. Or even sporting time everyday may be a useful independent variable, but these information is not included in the data set.

References

[1] "Health Insurance Data.", Kaggle
www.kaggle.com/bmarco/health-insurance-data.

[2] "About Adult BMI", Centers for Disease Control and Prevention (CDC)
https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

[3] "6 Reasons Healthcare Is So Expensive in the U.S.", Investopedia
https://www.investopedia.com/articles/personal-finance/080615
/6-reasons-healthcare-so-expensive-us.asp

[4] "Health Insurance Costs Surpass $20,000 Per Year, Hitting a Record", Bloomberg
https://www.bloomberg.com/news/articles/2019-09-25/
why-is-health-insurance-so-expensive-20-000-a-year-for-coverage