The Influence of Factors on Health Insurance

Research Question: In what ways do different factors influence insurance costs?

By: Crystal Chang, Jocelyne Rachelle Devera, Hiral Mehta, Evelina Ravlo, Isaac Yang

Introduction:

This dataset contains information on the relationship between medical insurance charges, personal attributes, and geographic factors. As healthcare insurance is a very important matter for many Americans, our goal is to determine how each of these variables influences insurance costs. In doing so, we will not only be able to make predictions about fees for different groups, but we will also be able to determine the probability of being insured for various population groups. When looking at this dataset, it is important to realize how healthcare insurance affects access to healthcare services, thereby affecting the health and economic well-being of the many people who make up the United States.

Data:

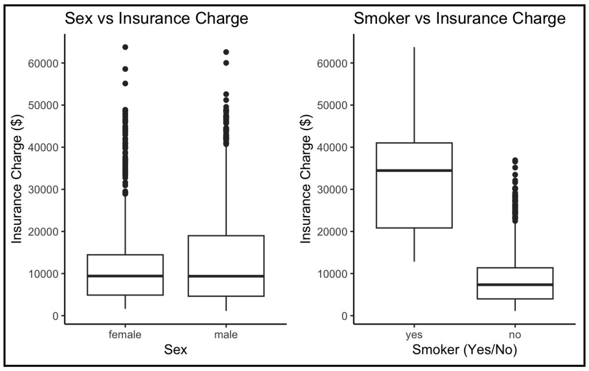
a. Data Description Table

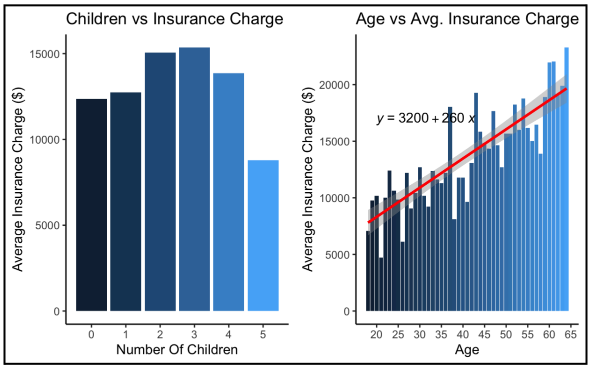
# of observations: 1338 observations

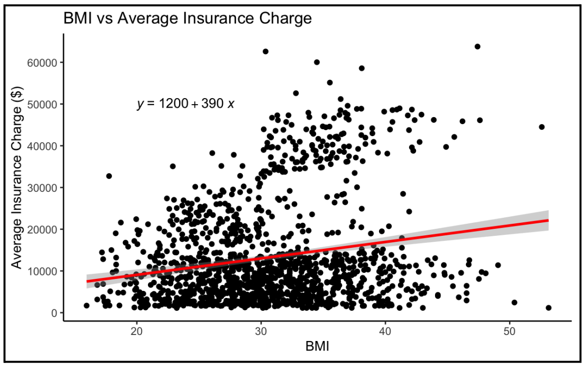
# of variables: 7 variables (1 dependent variable)

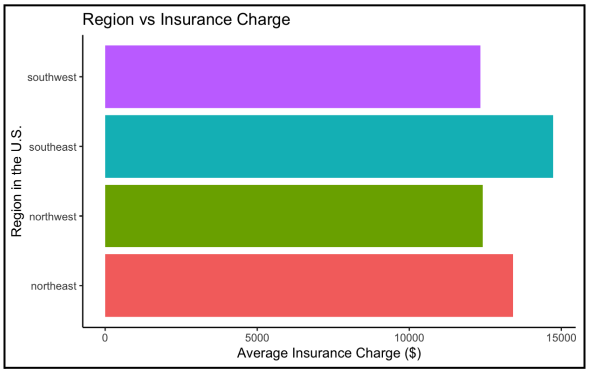
Variable	Description	Type
Insurance Charge β_0	*Dependent Variable* (Average insurance charge)	Numerical
Age X_1	*Independent* (Age in years)	Numerical
Sex X_2	*Independent* (Male vs Female)	Categorical
BMI X_3	*Independent* (Body Mass Index in kg/m^2)	Numerical
Children X_4	*Independent* (People with 1, 2, 3, 4, and 5 children)	Numerical
Smoker X_5	*Independent* (Smokers vs Non-smokers)	Categorical
Region X_6	*Independent* (East coast vs West coast)	Categorical

b. Graphs







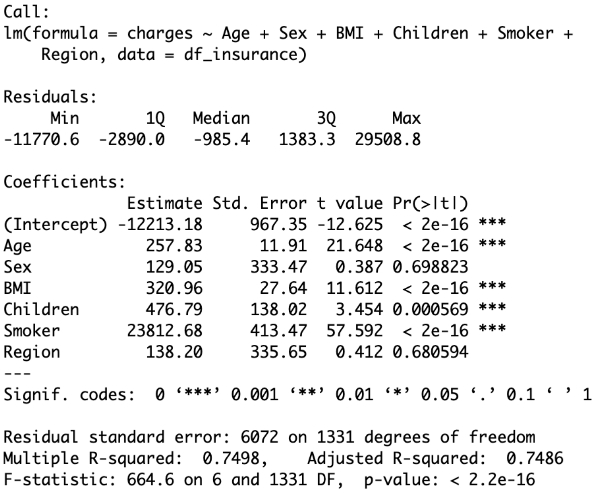


Results:

- **Proposed model (multivariate linear regression):**

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$

- **R-output (coefficient table):**

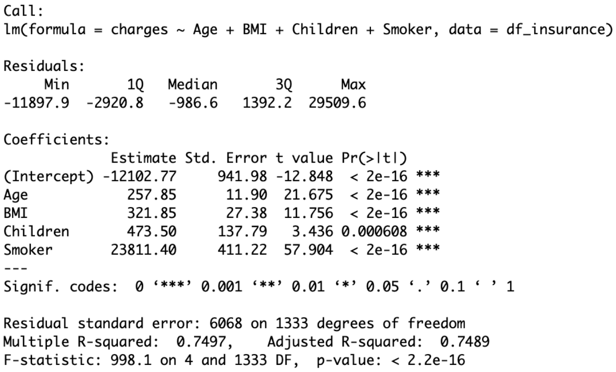


```
Call:
lm(formula = charges ~ Age + Sex + BMI + Children + Smoker + 
    Region, data = df_insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-11770.6  -2890.0   -985.4   1383.3  29508.8 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12213.18     967.35 -12.625  < 2e-16 ***
Age            257.83      11.91  21.648  < 2e-16 ***
Sex            129.05     333.47   0.387 0.698823    
BMI            320.96      27.64  11.612  < 2e-16 ***
Children       476.79     138.02   3.454 0.000569 ***
Smoker       23812.68     413.47  57.592  < 2e-16 ***
Region         138.20     335.65   0.412 0.680594    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6072 on 1331 degrees of freedom
Multiple R-squared:  0.7498,	Adjusted R-squared:  0.7486 
F-statistic: 664.6 on 6 and 1331 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = charges ~ Age + BMI + Children + Smoker, data = df_insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-11897.9  -2920.8   -986.6   1392.2  29509.6 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
Age            257.85      11.90  21.675  < 2e-16 ***
BMI            321.85      27.38  11.756  < 2e-16 ***
Children       473.50     137.79   3.436 0.000608 ***
Smoker       23811.40     411.22  57.904  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6068 on 1333 degrees of freedom
Multiple R-squared:  0.7497,	Adjusted R-squared:  0.7489 
F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Fitted model:

$$\widehat{y} = -12213.18 + 257.83X_1 + 320.96X_3 + 476.79X_4 + 23812.68X_5 + \epsilon$$

Interpretation of Graphs:

The sex vs average insurance charges box plots show that males and females tend to have similar average insurance charges. However, there are several outliers for these box plots, and the males have a much larger spread. The non-smoker box plot is normally distributed, whereas the smoker box plot appears to be slightly skewed. The median insurance costs for smokers and nonsmokers were roughly $35,000 and $7,500 respectively. Despite this compelling disparity, it might not be accurate as there are a lot of outliers in the nonsmoker box plot and there the smoker box plot has a lot more variation. For the bar graph with the number of children in a household, the highest insurance charges were among households with 2 or 3 children, whereas the lowest insurance charges were among households with 5 children. For the average American age vs Insurance charge graph, as age increases, the average insurance charge increases by about $260. The highest average insurance charge in this dataset was larger than $20,000 and was attributed to a 65-year-old individual. A weak, positive linear relationship can be observed in the scatterplot between the BMI and insurance charges because, for every increase in BMI by 1 kg/m^2, insurance charges increase by $390. All the regions have similar average insurance charges in the bar graph and as such, there does not appear to be much of a relationship between regions and insurance charges.

Interpretation of the Linear Regression:

Age, BMI, children, and smoking all have P-values less than 0.05, implying significance in relation to insurance charges. 74.97% of the variance within our data is explained by the regression line. There is also a positive relationship between age and charges, sex and charges, BMI and charges, children and charges, and smokers and charges. For every year added in age, the insurance charge increases by $273.83. For every unit of BMI, the insurance charge increases by $320.96. For every child added within a household, the insurance charge increases by $476.79. If one smokes, the insurance charge increases by $23,812.68.

Conclusion:

Higher vulnerability to diseases, increased fragility of bones and muscles, forgetfulness, and many other conditions stem from old age and are factors that can impact one’s health ([WHO], 2022). With so many health risks, it is reasonable to assume that medical insurance charges would be higher for older Americans.

Additionally, smoking can also lead to debilitating health problems like lung disease, diabetes, and cancer, all of which could lead to greater insurance costs ([CDC], 2020). When considering these health complications and their trends though, it is important to also recognize that the dataset may not be representative of the entire population. One particular factor to consider is the surveying method, as different methods can vary in their susceptibility to response bias. Participants who may be paying lower insurance charges might be inclined to respond that they do not smoke, or participants paying higher insurance charges may be inclined to say they smoke to get benefits from smoker rehab programs.

Comparing the relationship between the average number of children per household and insurance charges, households with 5 children do not have alarmingly high insurance charges. One possible explanation for this is the financial support and accommodations they are usually provided. With resources like health insurance to Medicaid/CHIP-eligible uninsured children, children of low-income households may improve their health at a lower out-of-pocket cost ([NIH], 2017). As a result, households with more children may have a lower recorded insurance charge than households with fewer children depending on the financial support threshold.

Increases in BMI can lead to possible increases in insurance charges. Obesity is a rising issue within the country, as it can lead to multiple health problems ([NIH], 2011). To remedy this, higher insurance charges may be implemented. Then again, there are other possible reasons for the increase in insurance charges aside from BMI because the depicted regression line is not that strong.

References:

- https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance
- https://www.who.int/news-room/fact-sheets/detail/ageing-and-health#:~:text=Common%20conditions%20in%20older%20age,%2C%20diabetes%2C%20depression%20and%20dementia.
- https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm#:~:text=Smoking%20causes%20cancer%2C%20heart%20disease,immune%20system%2C%20including%20rheumatoid%20arthritis.
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5463460/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6415902/