MA678 Midterm Project Report

Yingjie Wang

Abstract

Health insurace is an important part of daily life, and its charges depend on a lot of factors. In this project, a health insurance datast is analyzed with the goal of determining what factors affect the amount of health insurance charges. To address this problem, a generalized linear regression model is implemented to reveal related factors.

Introduction

Medical cost is one of the most expensive cost in daily life, especially for elder people. It is thus an important question that what factors affect the health insurance charges. In "Medical Cost Personal Datasets" (https://www.kaggle.com/mirichoi0218/insurance) on Keggle platform, there are 1338 samples of healthy insurance charges data. This dataset consists of 7 attributes: age, sex, bmi, children, smoker, region and charges. Among all the attributes, age, bmi, children and charges are numerical attributes, while sex, smoker and region are categorical attributes.

The goal of this project is to understand what factors affect the amount of health insurance charges, therefore the outcome of the regression model is attribute charges, and all other attributes form the predictors of the regression model.

Methods

Data Preparation

For preparing the dataset, I loaded the dataset "insurance.csv" in RStudio. The dimension of this dataset is (1338, 7), which means this dataset has 1338 samples and 7 attributes. Here are some explanation of columns:

column names	explanation
age	Age of primary beneficiary
sex	Insurance contractor gender (female/male)
bmi	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height
children	Number of children covered by health insurance / Number of dependents
smoker	Binary indicator of whether the insurance contractor is a smoker
region	The beneficiary's residential area in the US, (northeast, southeast, southwest, northwest).
charges	Individual medical costs billed by health insurance

Exploratory Data Analysis

First, I explored the distribution of attributed. For numerical attributed, histograms are created:

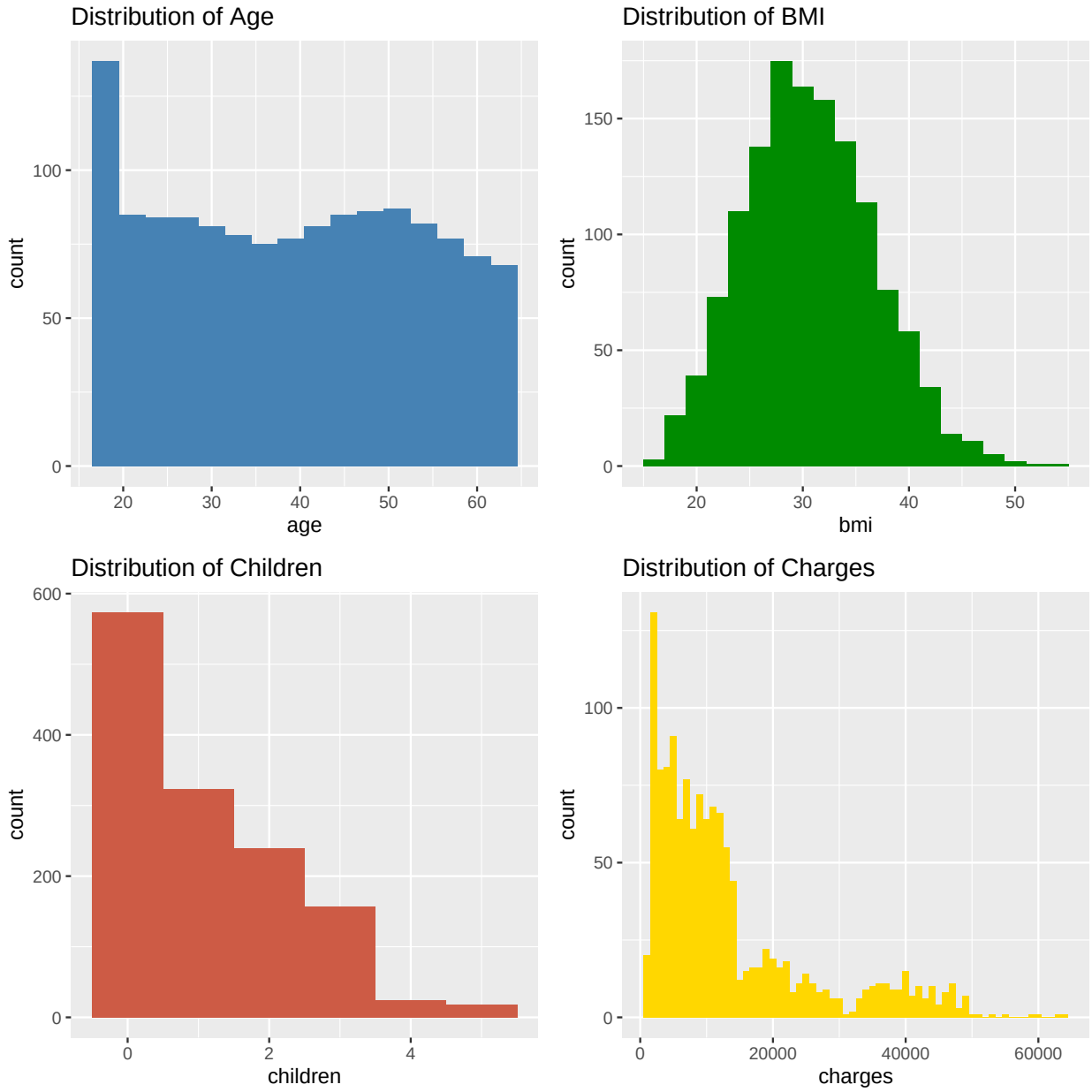

Figure 1: Histograms of numerical attributes.

Considering age is very likely to be related to health insurance charges, scatter plots of the relationship between attribute "age" and "charges" are plotted. There are some factors that may influence the relationship between "age" and "charges". Among all other attributes, "smoker" amd "sex" are selected.

From these two scatter plots, relationship between the age and charges are highly influenced by whether the insurance contractor is a smoker, with a smoker contractor paying 24000 more than a non-smoker on average for health insurance. However, there is no obvious difference when grouping by sex attribute, with male contractor pays slightly more than female contractor on average.

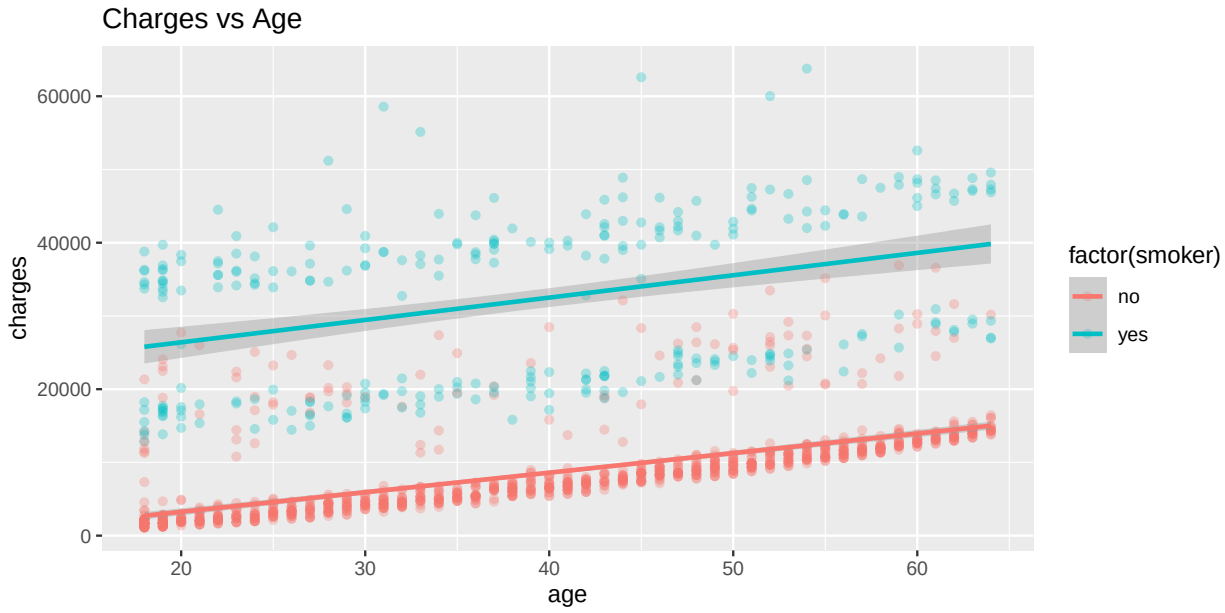

Figure 2: Scatter plot showing the relationship between age and charges, grouped by whether the insurance contractor is a smoker.

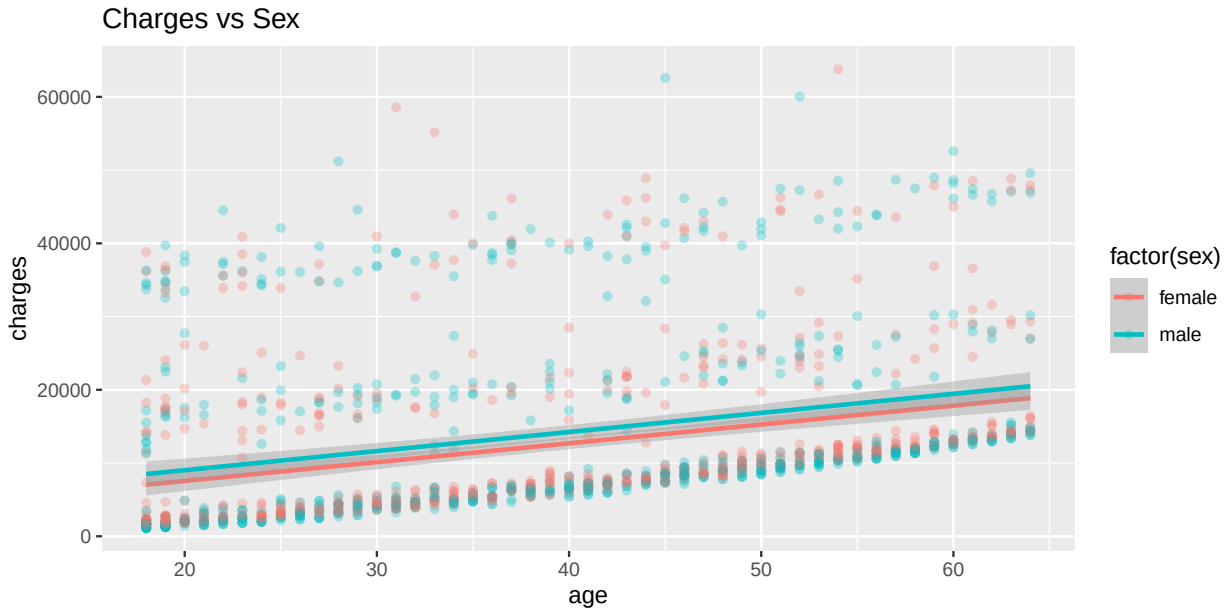

Figure 3: Scatter plot showing the relationship between age and charges, grouped by the gender of the insurance contractor.

Model Fitting

A generalized linear model is implemented to capture the relationship between the health insurance charges and predictors. The attribute "region" is not included in the predictors, since intuitively, the health insurance is not likely to be influenced by general region. Considering whether a contractor is a smoker is likely to be related to the gender of the contractor, the interaction between attribute "smoker" and attribute "sex" is included. BMI is a measure related to health, and very-low or very-high BMI are both likely to reduce the health level of the contractor, therefore a quadratic term of BMI is also included in the regression model.

Below is the function:

```
model <- glm(charges ~ age + bmi + I(bmi^2) + sex * smoker + children, data=insu)
```

Result

Model Coefficients

The formula of our fitted model is shown below:

$$charges = -18177.710 + 256.317 \cdot age + 748.177 \cdot bmi - 6.809 \cdot bmi^2 - 568.723 \cdot I_{male} + 22609.172 \cdot I_{smoker} + 464.322 \cdot children + 2200.165 \cdot I_{male} \cdot I_{smoker}$$

Model Validation

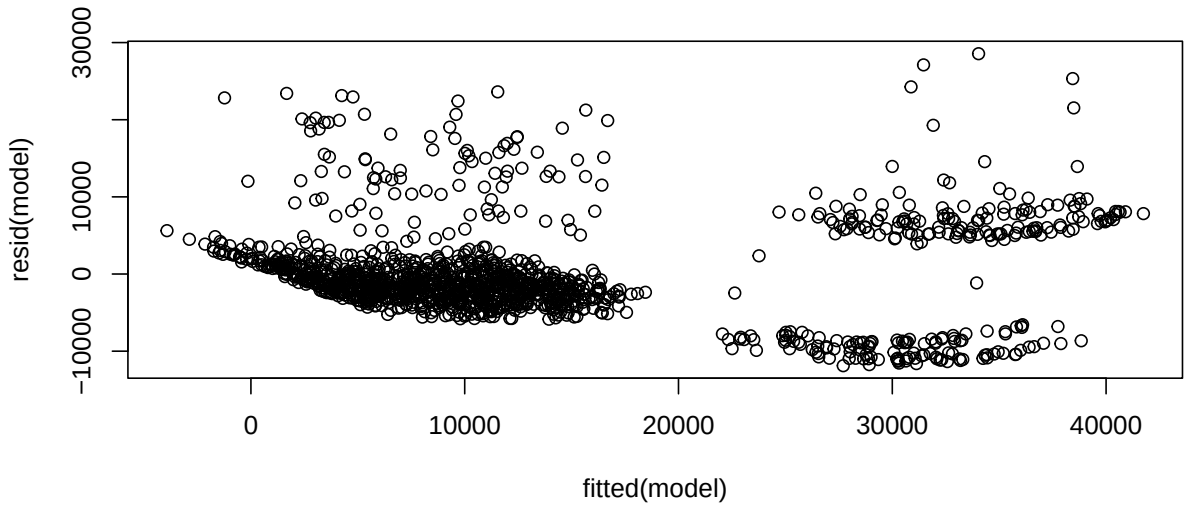

Figure 4: Residual plot.

From the residual plot, there are some outliers. The possible reason is that there exist some factors that influence the health insurance that are not included in the dataset. For example, some insurance contractors may have some disease history that increase the health insurance charges, however, it is not shown in the dataset. This can well explain the outliers on the upper half of the residual plot.