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 dataset is analyzed with the goal of determining what factors affect the amount of health insurance charges. Considering non-smoker group and smoker group can influence the way predictors affect the outcome, a multilevel model is implemented to reveal related factors. The model shows that there is clear difference between the way predictors affect the outcome for non-smoker group and smoker group, both for intercept and the slope of age variable.

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).

column names	explanation
charges	Individual medical costs billed by health insurance

Exploratory Data Analysis

First, I explored the distribution of numerical attributes and the overall histograms are shown in appendix. Among all numerical attributes' distributions, charges' distribution is far from normal distribution, so I took the log of this attribute where the transformed distribution is more like a normal distribution. The two distributions (before and after log transformation) are shown in Figure 1. For all the downstream analysis, log transformation is used for the charges attribute.

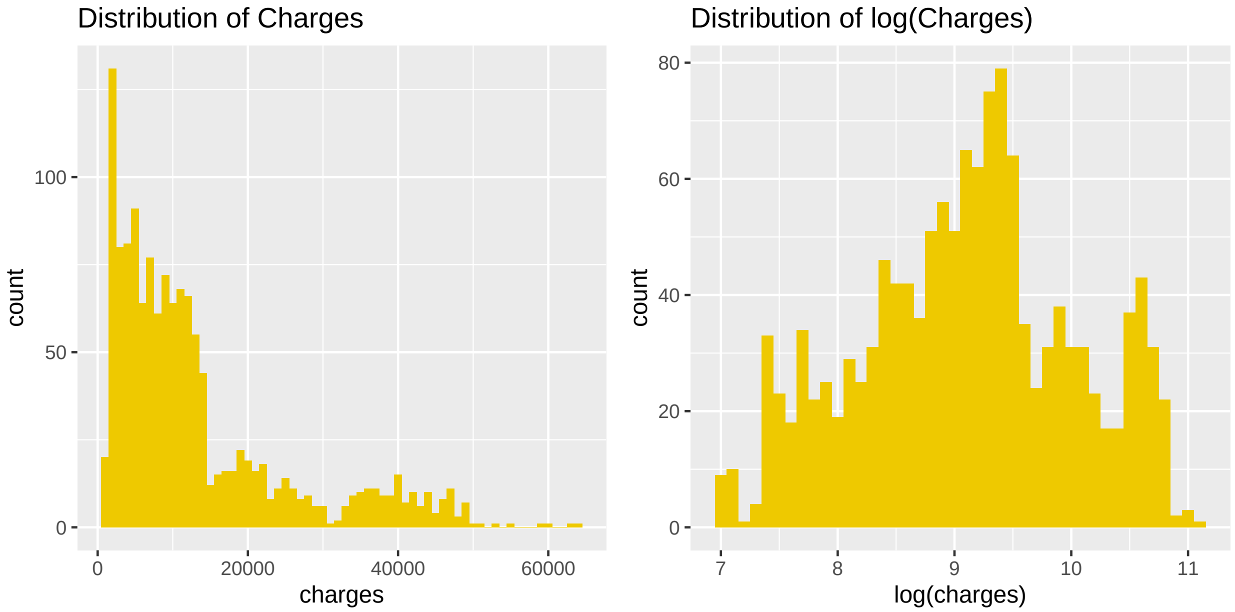

Figure 1: Histograms of numerical attributes.

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

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

Model Fitting

A multilevel model is used to fit the data considering different categories. Given the distribution of "charges" attribute, a log transformation is used. Therefore, log(charges) is the outcome of the model. Since from EDA it is noted that the smoker categories (whether an insurance contractor is a smoker) have different correlation with variables, I used varying slope and varying intercept in this multilevel models.

Below is the function:

```r
model <- lmer(log(charges) ~ age + sex + bmi + children + (1 + age|smoker), data=insu)
```

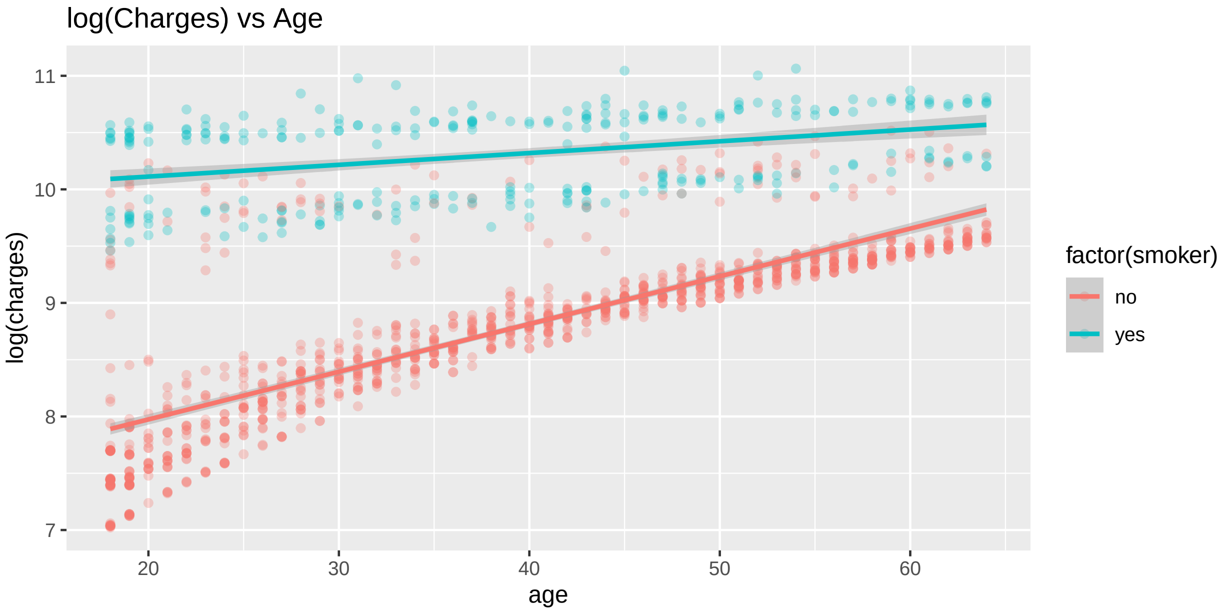

Figure 2: Scatter plot showing the relationship between age and log(charges), grouped by whether the insurance contractor is a smoker.

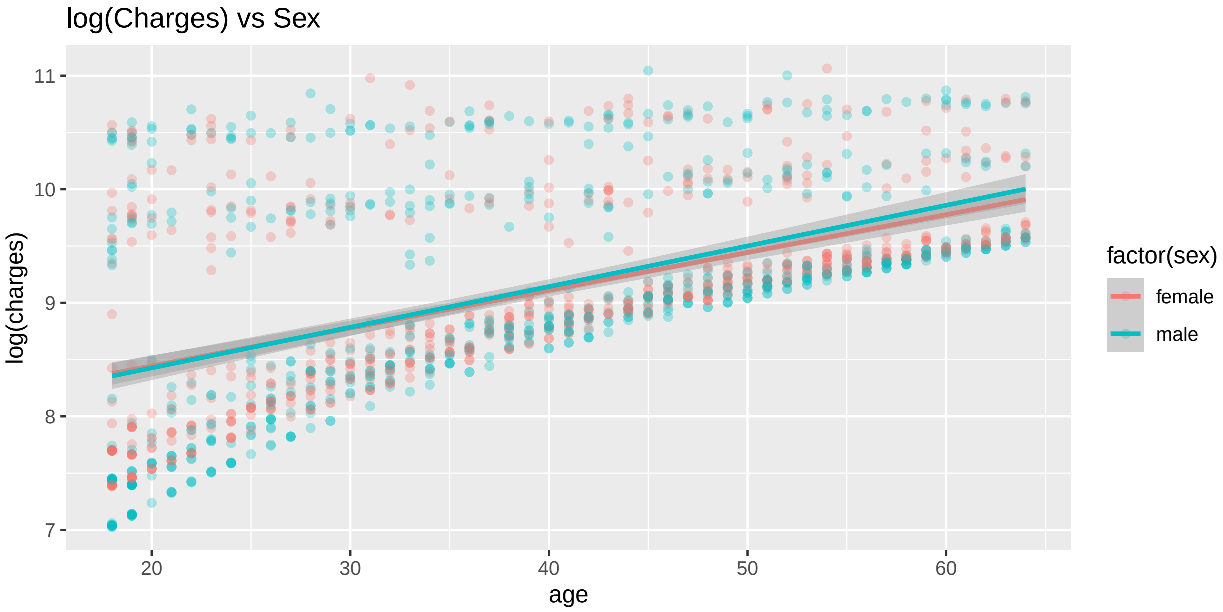

Figure 3: Scatter plot showing the relationship between age and log(charges), grouped by the gender of the insurance contractor.

Result

Model Coefficients

The fixed effects of our fitted model is shown below:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	8.165	7.622e-01	4.544	10.712	0.000214 ***
age	2.529e-02	8.691e-03	4.458	2.91	0.038306 *
sexmale	-7.228e-02	2.255e-02	1331	-3.205	0.001384 **
bmi	1.02e-02	1.856e-03	1331	5.496	4.65e-08 ***
children	1.044e-01	9.330e-03	1331	11.189	< 2e-16 ***

From the fixed effect table, the attribute age is not significant at alpha = 0.05 level. This might be caused by the difference of age attribute's influences on the outcome.

For different categories, each predictor always has different influence on the outcome. Below is random effects table for different categories:

	(Intercept)	age	sexmale	bmi	children
Non-smoker	6.7794	0.0411	-0.0723	0.0102	0.1044
Smoker	9.5497	0.0095	-0.0723	0.0102	0.1044

Therefore, for non-smoker, we can conclude the formula as follow:

$$log(charges) = 6.7794 + 0.0411 \cdot age - 0.0723 \cdot I_{male} + 0.0102 \cdot bmi + 0.1044 \cdot children$$

For smoker, we have another formula:

$$log(charges) = 9.5497 + 0.0095 \cdot age - 0.0723 \cdot I_{male} + 0.0102 \cdot bmi + 0.1044 \cdot children$$

Model Validation

A residual plot is shown in Appendix Figure 4. 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.

Discussion

In this project, predictors' relationship with the outcome log(charges) is analyzed. Due to helpful EDA, it is noted that the distribution of attribute "charges" is far from normal distribution which leads to a log-transformation. It is also important to note that attribute "age" have different influence to the outcome for different smoker categories (non-smoker and smoker). Considering these two facts, a multilevel model with outcome "log(charges)" is implemented. The result of this model highly agree with out assumption that the "smoker" category can affect the way attribute "age" influence the outcome.

Appendix

EDA histograms

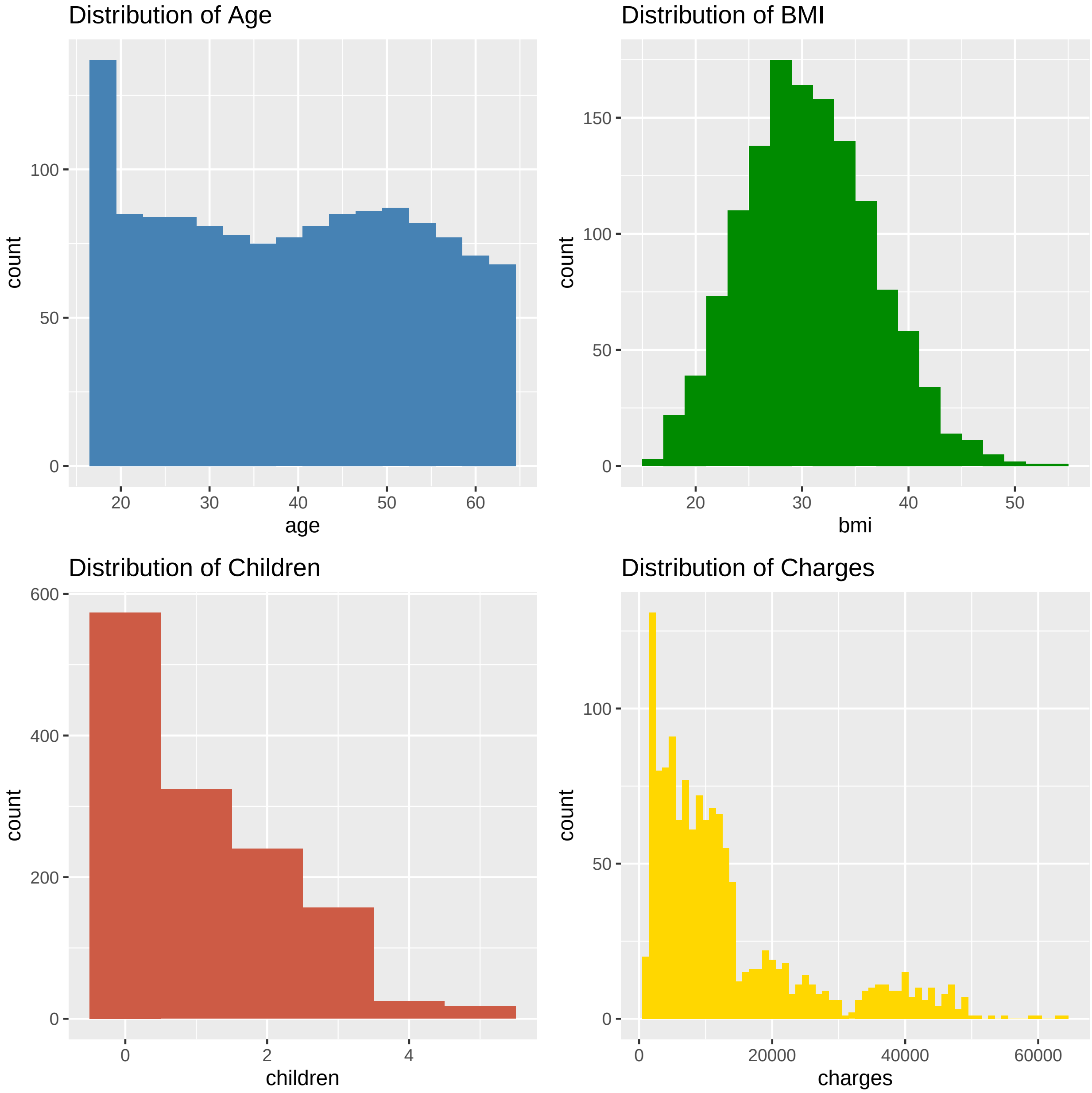

Figure 4: Histograms of numerical attributes.

Residual plot

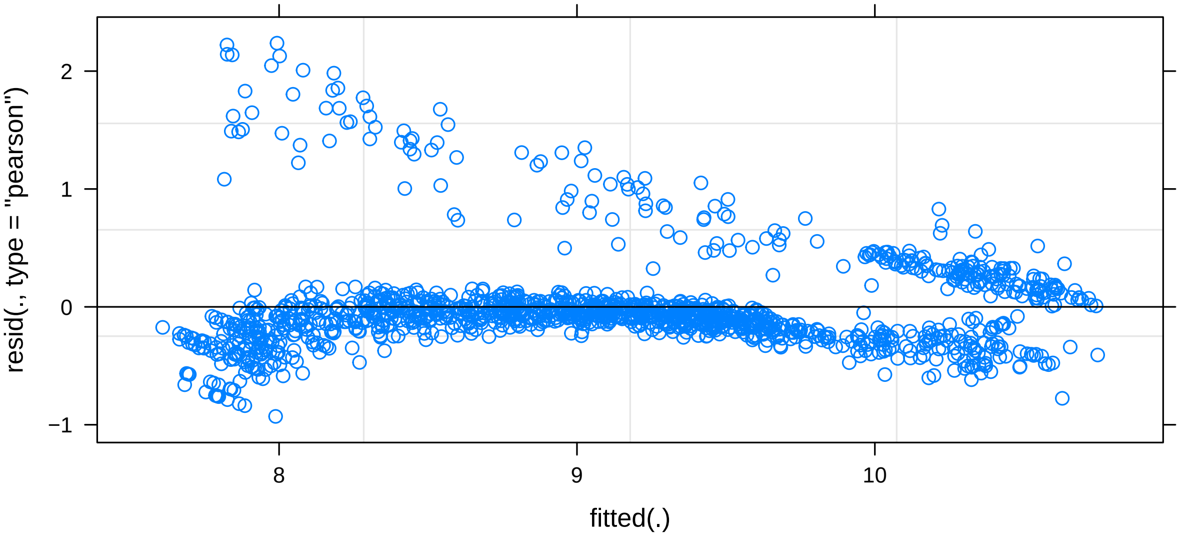

Figure 5: Residual plot.