MA678 Midterm Report

-Health Insurance Charges

Xiang Li

11/30/2021

Abstract

Health insurance pays for some or all the cost of the health services you receive, like doctors' visits, hospital stays, and visits to the emergency room. It helps keep your health care costs predictable and affordable. Although being covered by your insurance, you may still have to pay amounts for health insurance, such as premium. In this report, I will analyze how the insurance charges are affected by other factors.

Introduction

Health insurance pays for some or all the cost of the health services you receive, like doctors' visits, hospital stays, and visits to the emergency room. It helps keep your health care costs predictable and affordable. Although being covered by your insurance, you may still have to pay amounts for health insurance, such as premium. In this report, I will analyze how the insurance charges are affected by other factors.

Method

Data Wrangling

Datasource used in this report is a dataset named *US Health Insurance Dataset* from Kaggle. This dataset is a mix of numeric and categorical variables. There are seven variables and 1337 observations, where the Insurance charges are given against the following attributes of the insured: Age, Sex, BMI, Number of Children, Smoker, and Region. According to CDC (https://www.cdc.gov/healthyweight/assessing/index.html), I divided BMI data into 4 groups and age data into 6 groups preparing for the following EDA.

Column names	Explanation
age	Age of primary beneficiary
sex	Insurance contractor gender, female / male
bmi	Body mass index
children	Number of children covered by health insurance
smoker	Smoker / Non - smoker
region	The beneficiary's residential area in the US, northeast/southeast/ southwest/northwest
charges	Individual medical costs billed by health insurance.

BMI Data	BMI Group	AGE Data	AGE Group
bmi<=18.5	UnderWeight	age <=20	Group1
18.5< bmi<=24.9	HealthyWeight	20< age <=30	Group2
25<bmi<=29.9	OverWeight	30< age<=40	Group3
bmi>=30	Obese	40< age<=50	Group4
		50 < age<=60	Group5
		age> 60	Group6

Exploratory Data Analysis

At first, I generated a correlation matrix to give me a basic sense of correlation between each factors.

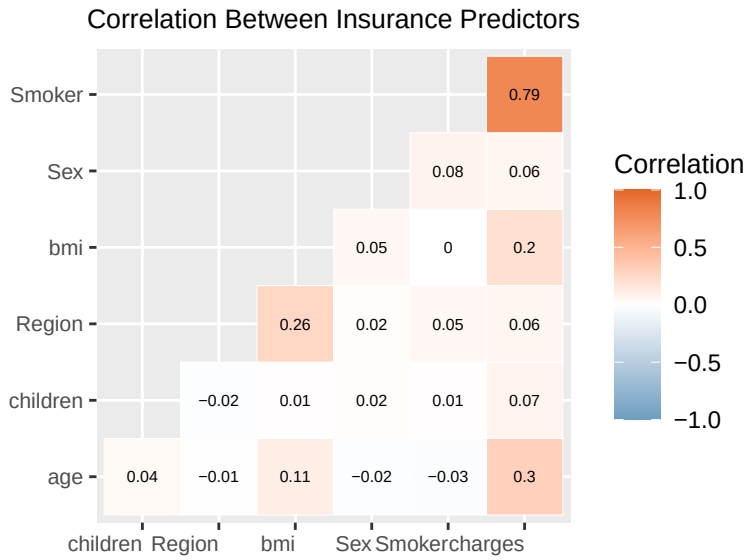
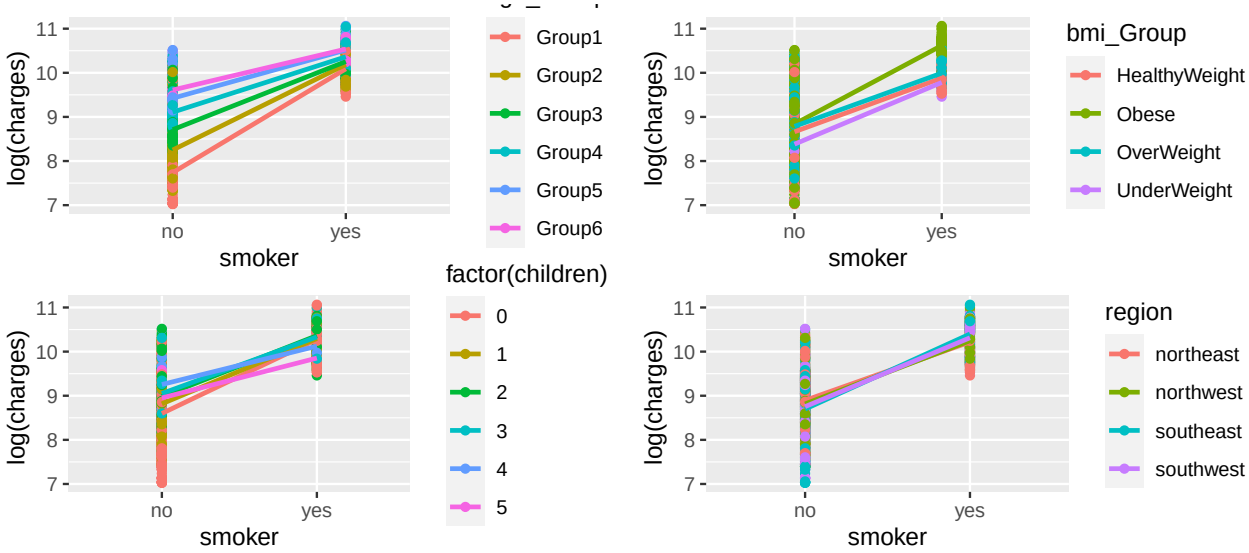
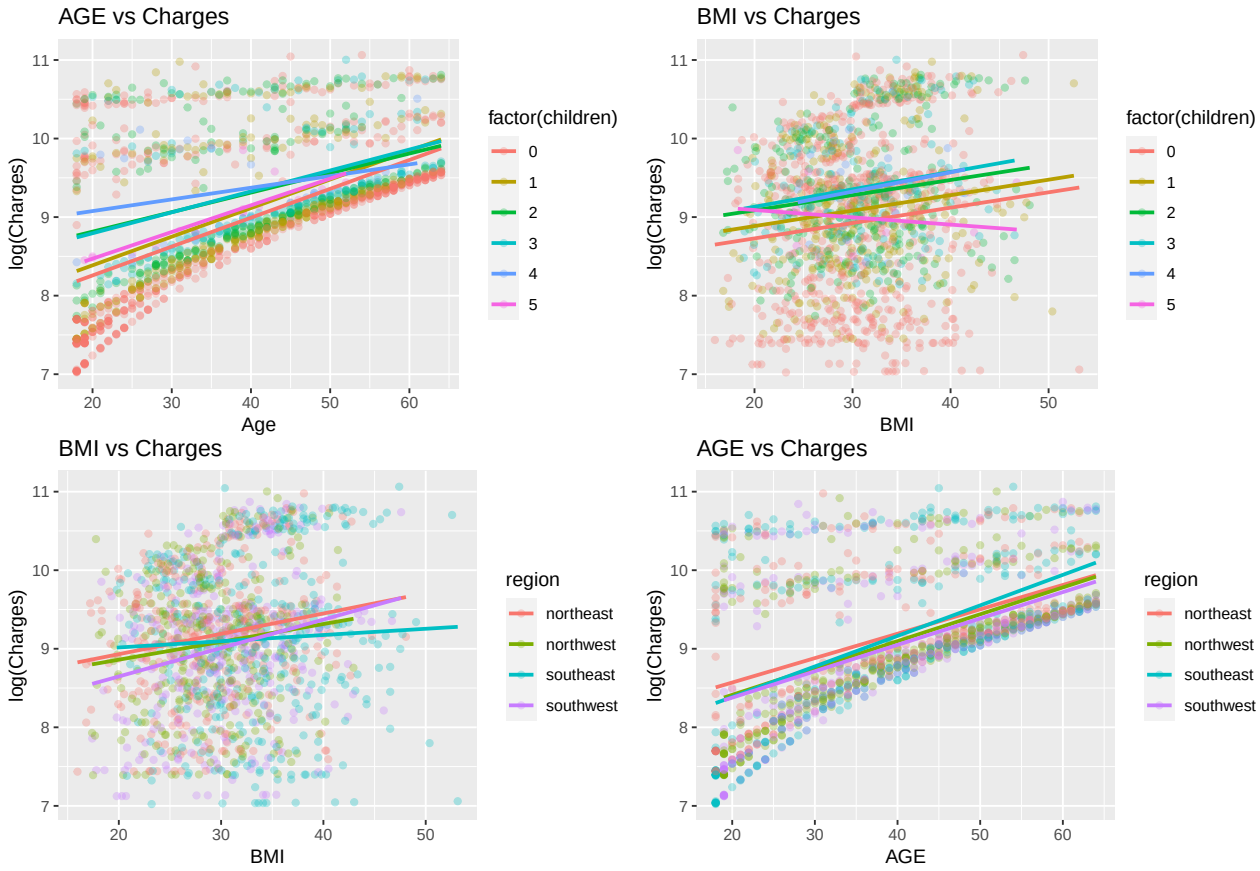


Figure 1: Correlation plot

According to the previous correlation plot, I can conclude that the premium charges show a strong positive correlation with smoking habits. Therefore, I'd like to analyze when outcome is log(charges) the relationship between smoking habits and several random effects which are age group, BMI group, children, and region.






```
##                   Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)         7.5721     0.0946   10.2401 80.0428   0.0000
## age                 0.0346     0.0014    4.3461 25.5859   0.0000
## sexmale            -0.0843     0.0207 1311.6048 -4.0677   0.0001
## smokeryes           1.1669     0.3039    8.4106  3.8401   0.0045
## children            0.0566     0.0136    9.1518  4.1667   0.0023
## regionnorthwest    -0.0620     0.0297 1312.9264 -2.0872   0.0371
## regionsoutheast    -0.1338     0.0292 1152.2136 -4.5881   0.0000
## regionsouthwest    -0.1526     0.0298 1308.0456 -5.1269   0.0000
```

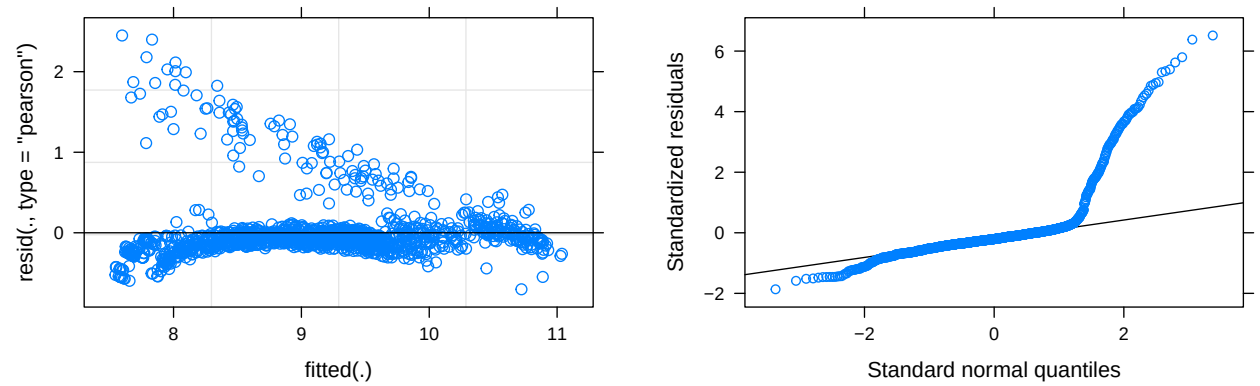


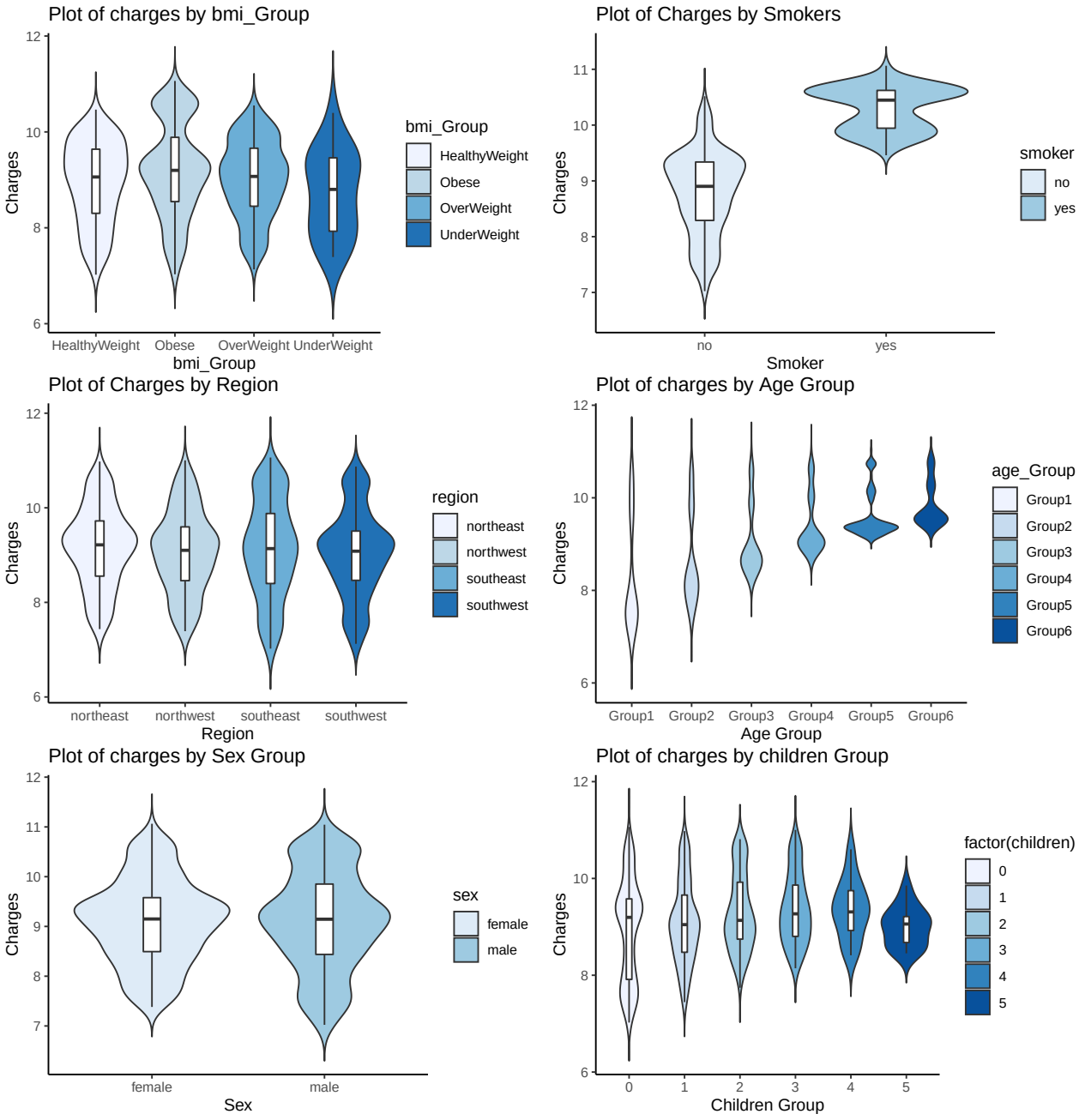
Figure 2: Residual plot and Q-Q plot

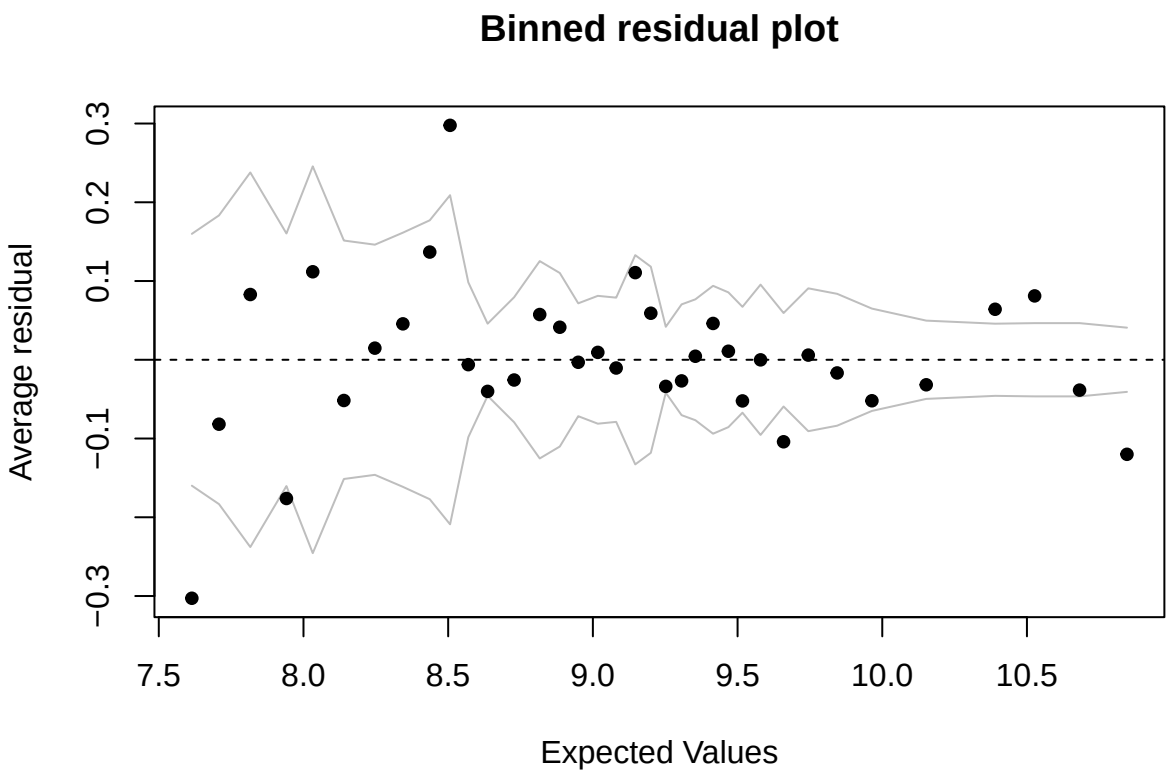
Result

Discussion

Appendeix

In appendeix, I will show more EDA that I generated.
Here are violin plots showing the distributions of chargers and each of other factors.




Binned residual plot
Average residual
0.3
0.2
0.1
−0.1
−0.3
7.5
8.0
8.5
9.0
9.5
10.0
10.5
Expected Values