Health Insurance Charges

Xiang Li

11/30/2021

Abstract

Health insurance provides financial protection in case people have a serious accident or illness. Health care can be very expensive. It can be an enormous financial burden. Therefore, health insurance is important to have health insurance as a safety net. However, what will make insurance company charges are so different. In other words, which people choose expensive insurance packages and which group of people choose cheaper insurance packages. I explored the US health insurance data set from Kaggle which included 1337 observations to analyze how the insurance `charges` are affected by other factors such as `age`, `sex`, and so on. By using the multilevel linear regression model, I conduct that `smoking habit` has the most significant effect on the insurance charges and `age` and `bmi` have a slight positive effect on the insurance charges. I use three groups as a random effect, which are `age_group`, `bmi group`, and `region` and state that variables mentioned before are slightly different between random effects. In this report, there are four main parts, which are Introduction, Method, Result, and Discussion.

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. What kind of package people choose decides the coverage by the insurance company. Data source 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`.

Method

According to CDC's weight assessment, I divided BMI data into 4 groups and age data into 6 groups preparing for the following EDA.

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

In the beginning, I make density plots to see whether the charge is following a normal distribution. Using `bmi_Group` as an example to see the distribution.

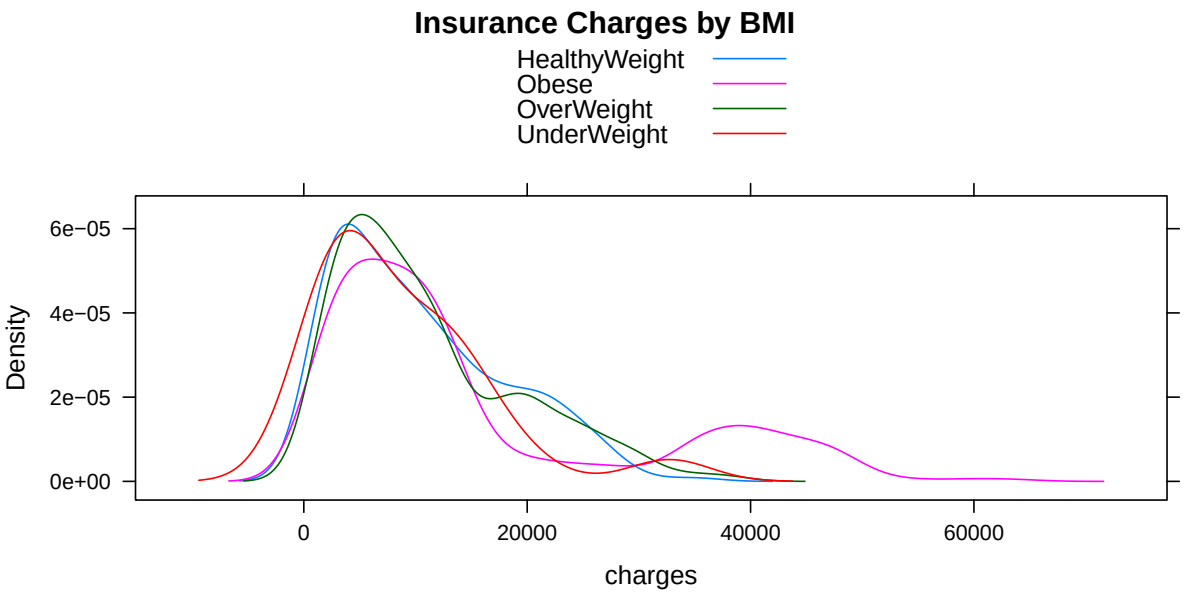


Figure 1: Distribution of charges among different bmi group

Since figure 1 showed a right-skewed distribution, I use `insurance charges` in log transformation in the following EDA and model fitting.

I'd like to analyze when the outcome is `log(charges)`, the relationship between smoking habits, and several random effects which are `age_Group`, `bmi_Group`, `children`, and `region`.

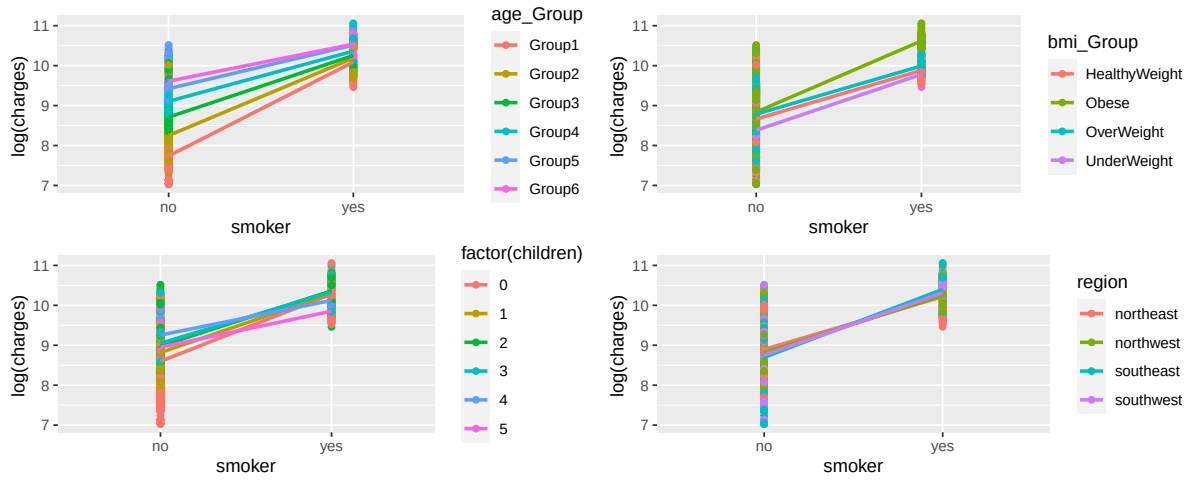


Figure 2: Relationship between smoker and charges

Figure 2 shows that `age_Group` has different effects on smoking habits with different interceptions and slopes. Since the older group has a smaller slope than the young group. For other random effects, there is a barely obvious distinction.

Then I'd like to analyze when the outcome is `log(charges)` the relationship between `sex` and several random effects which are `age_Group`, `bmi_Group`, `children`, and `region`.

Figure 3 shows that `bmi_Group` has different effects on `Sex`. Underweight and healthy-weight males have less charge than underweight and overweight females. However, overweight and obese males have

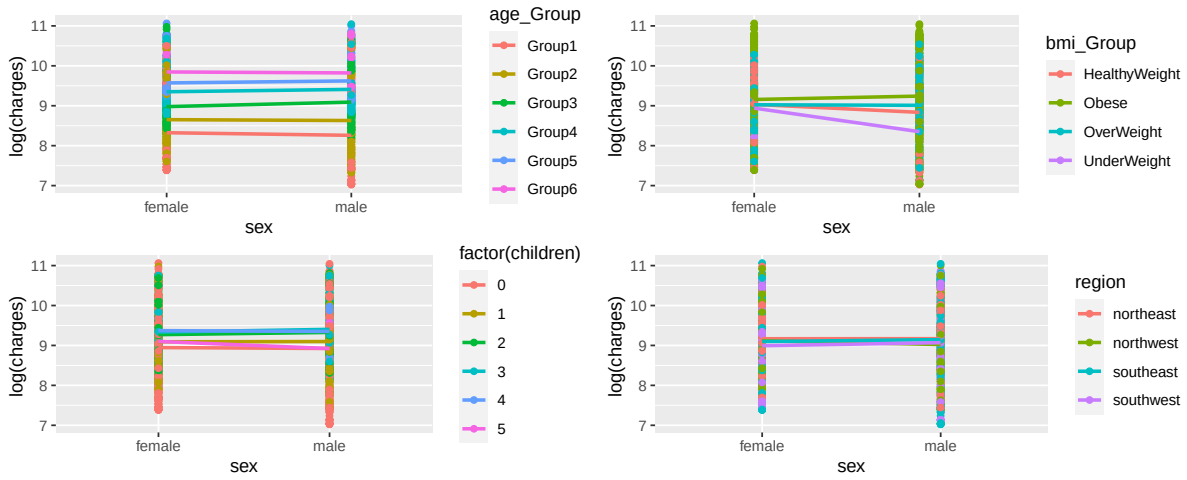


Figure 3: Relationship between sex and charges in four different group

less charge than Underweight and overweight females. For other random effects, they don't have a clear distinction to the variable of `Sex`.

Next, I'd like to analyze when the outcome is `log(charges)` the relationship between `age` and `insurance charges` & `bmi` and `insurance charges` when a random effect is `region`.

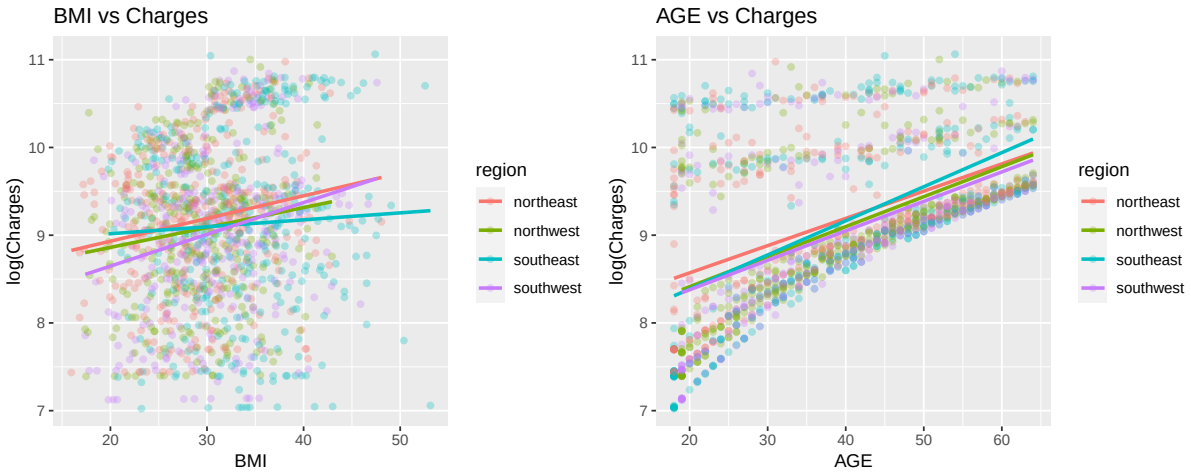


Figure 4: Relationship between age and insurance charges & bmi and insurance charges when random effect is region

Figure 4 shows that different region has different effects on age since I observed distinct slopes and different intercepts. In the `southeast`, the slope is obviously more placid than others. In the right figure, the random effect of `region` doesn't have a clear distinction to the variable of `age`. Therefore, the random effect of the region has slightly difference on `bmi`.

Model Fitting

In accordance with the previous EDA section, I decided to use three groups as random effects, which are `age_Group`, `bmi_Group` and `region`, since when they are as random effects, the following function is the model I build for my research.

```
fit_model <-  lmer(log_charges~Sex+age+bmi+Smoker+
                     (1+Smoker|age_Group)+
```

```
                    (1+Sex|bmi_Group)+
                    (1+bmi|region),data=insurance)
```

Here is the summary of all fixed effects and all variables are considered as statistically significant at $\alpha = 0.5$ level.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	6.9847	0.1491	16.3060	46.839	< 2e-16 ***
Sex	-0.0720	0.0240	13.0322	-3.000	0.01021 *
age	0.0386	0.0026	9.0887	14.804	1.14e-07 ***
bmi	0.0091	0.0036	7.9117	2.501	0.03720 *
Smoker	1.5278	0.2141	4.6960	7.136	0.00109 **

Therefore, the final model is

$$log(charges) = 6.9847 - 0.072 \cdot Sex + 0.0386 \cdot age + 0.0091 \cdot bmi$$
$$+1.5278 \cdot Smoker + ranef_{bmi_Group} + ranef_{age_Group} + ranef_{region}$$

After deciding which model I would use through ANOVA comparison, I check the binned residual plot, residual vs fitted plot, and QQ plot, which are listed in the Appendix in the end of the report. In the QQ plot, almost one-third part of the points do not lie approximately on a straight line, which indicates that possible outliers are not in a normal distribution, distance from the bulk of the observations. Based on the binned residual plot, the model looks reasonable.

Result

Model Coefficients

Here are the coefficients of the random effect of `age_Group`.

```
##        (Intercept)      Smoker
## Group1 -0.20297295  0.79932270
## Group2  0.04000267  0.39093426
## Group3  0.10192445  0.01483718
## Group4  0.09269817 -0.22953307
## Group5  0.03261265 -0.41378300
## Group6 -0.06426499 -0.56177808
```

Here are the coefficients of the random effect of `bmi_Group`.

```
##                (Intercept)          Sex
## HealthyWeight -0.01156509 -0.003085412
## Obese          0.04507626  0.012025473
## OverWeight    -0.01397429 -0.003728021
## UnderWeight   -0.01953687 -0.005212041
```

Here are the coefficients of the random effect of `region`.

```
##           (Intercept)          bmi
## northeast -0.06836931  0.004838307
## northwest -0.01829490  0.001294679
## southeast  0.04534403 -0.003208872
## southwest  0.04132017 -0.002924114
```

And let's take the insured who are obese from the southeastern part of the United State who is in their thirties as an example. I would like to conduct the following formula.

$$log(charges) = 7.177 - 0.06 \cdot Sex + 0.0386 \cdot age + 0.0544 \cdot bmi + 1.5426 \cdot Smoker$$

For every 1% growth in age, the predicted insurance charge of the insured who are obese from the southeastern part of the United State who is in their thirties will be increasing 5.44%. It is similar interpretations to other group people.

Discussion

From the previous processing, the result generated mostly are expected. People who have smoking habits are likely to purchase more expensive insurance packages and pay more premiums. And there is not surprising that the insured who has an older age, the premium will be higher. However, what I am not expecting is that underweight people have lower insurance charges than healthy-weight people. But it makes sense though, nowadays people do not pay much attention to the underweight group because common diseases such as hypertension and diabetes are obesity diseases. That means underweight people do not think they will experience chronic diseases, so they don't purchase very expensive insurance

Moreover, regarding the dataset itself, it has limitations since there are only 1337 observations with seven predictors, and also the dataset is in 2019, two years ago before the Covid-19 pandemic. I believe that some people's ideas about insurance will change significantly after the epidemic, for example, they will upgrade their insurance to obtain more coverage. In other words, for the insured who have kids under the insurance, if they want to upgrade the package, the growth will be multiplied and even exponential.

Since I am really interested in the topic, if there is updated data coming in, I will keep doing an analysis on the insurance charges.

Reference

US Health Insurance Dataset, Anirban Datta, https://www.kaggle.com/teertha/ushealthinsurancedataset

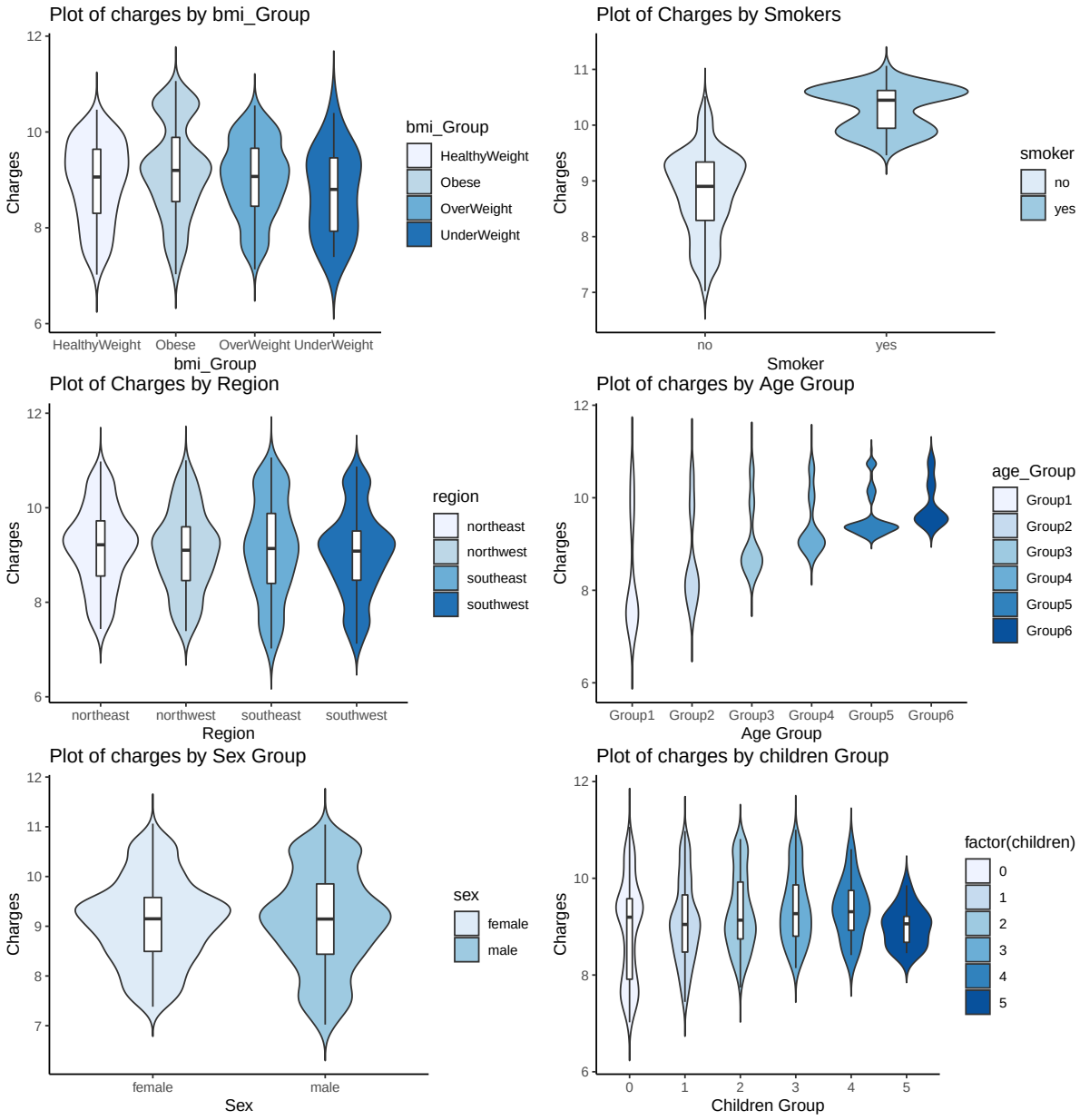
Centers for Disease Control and Prevention, Assessing Your Weight, https://www.cdc.gov/healthyweight/assessing/index.html

Appendix

This frame is the explanation of US Insurance Charges from Kaggle.

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.
Smoker	Yes=1,No=1

Here are violin plots showing the distributions of chargers and each of other factors.



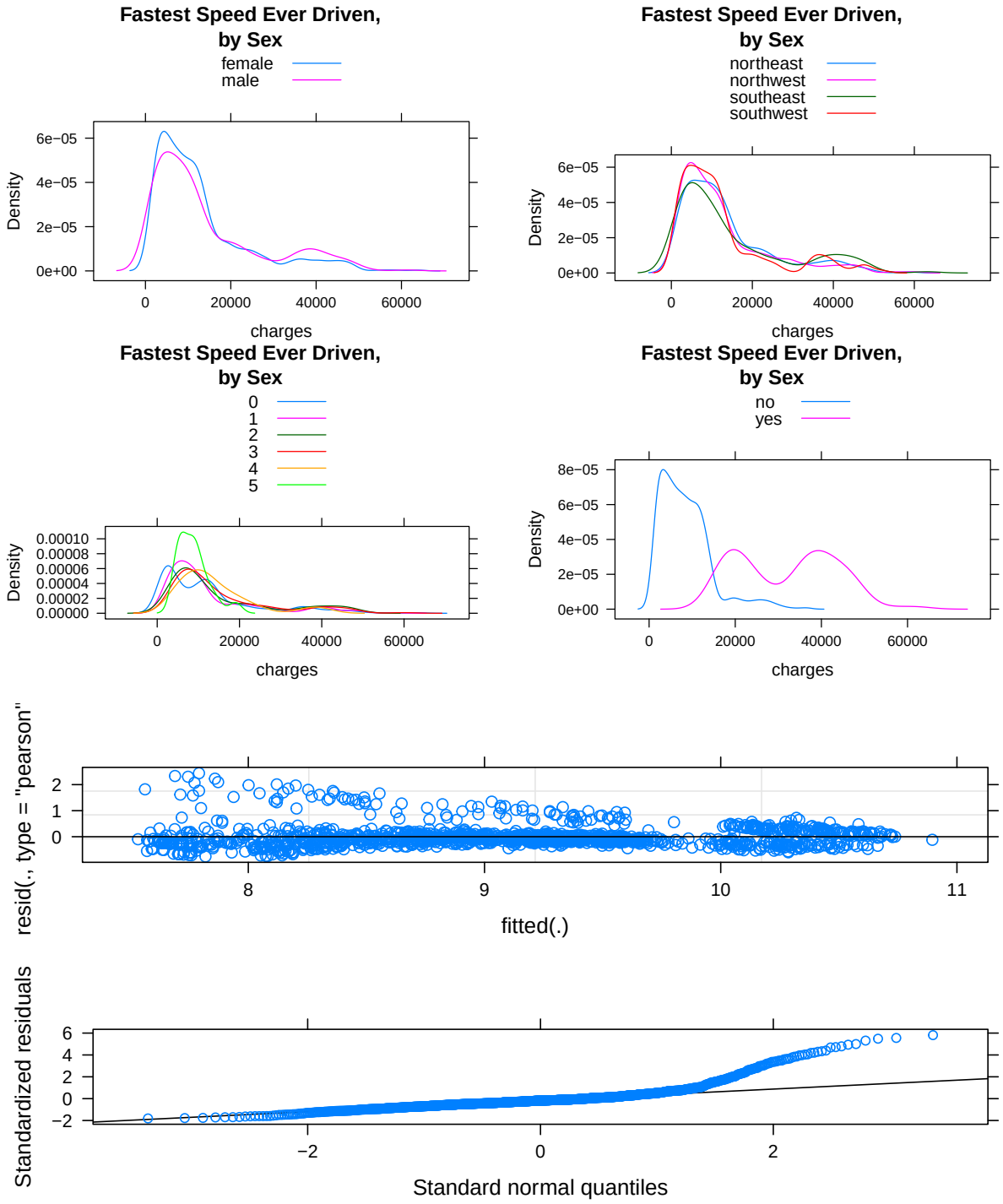


Figure 5: Residual plot and Q-Q plot

```
## $age_Group

##
## $bmi_Group

##
## $region
```

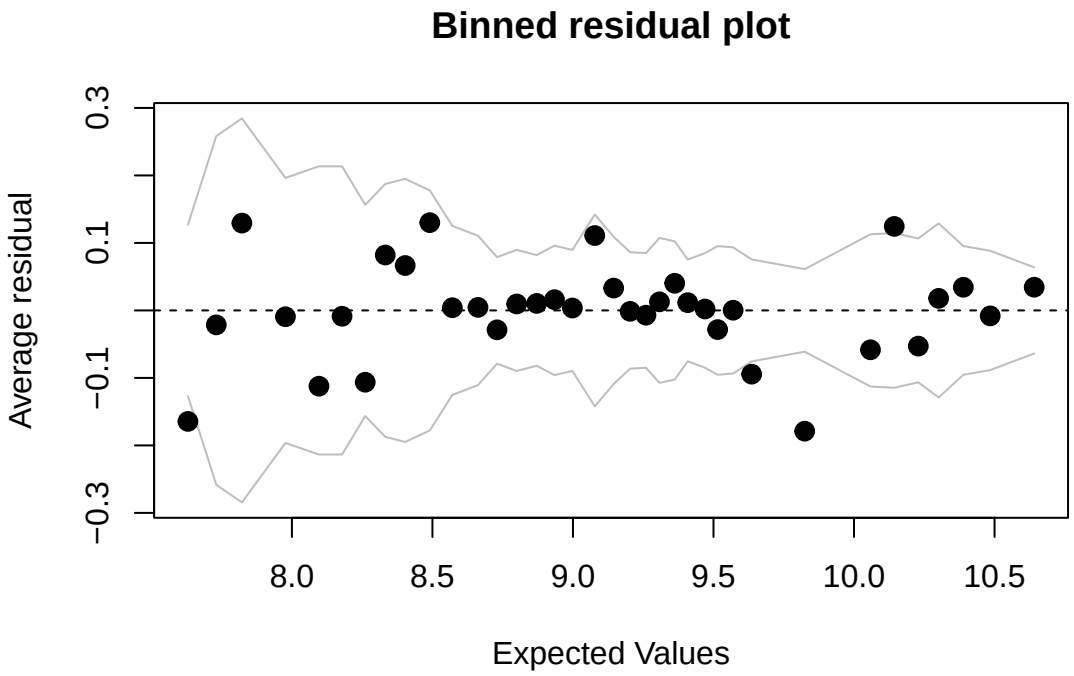


Figure 6: Residual binned plot

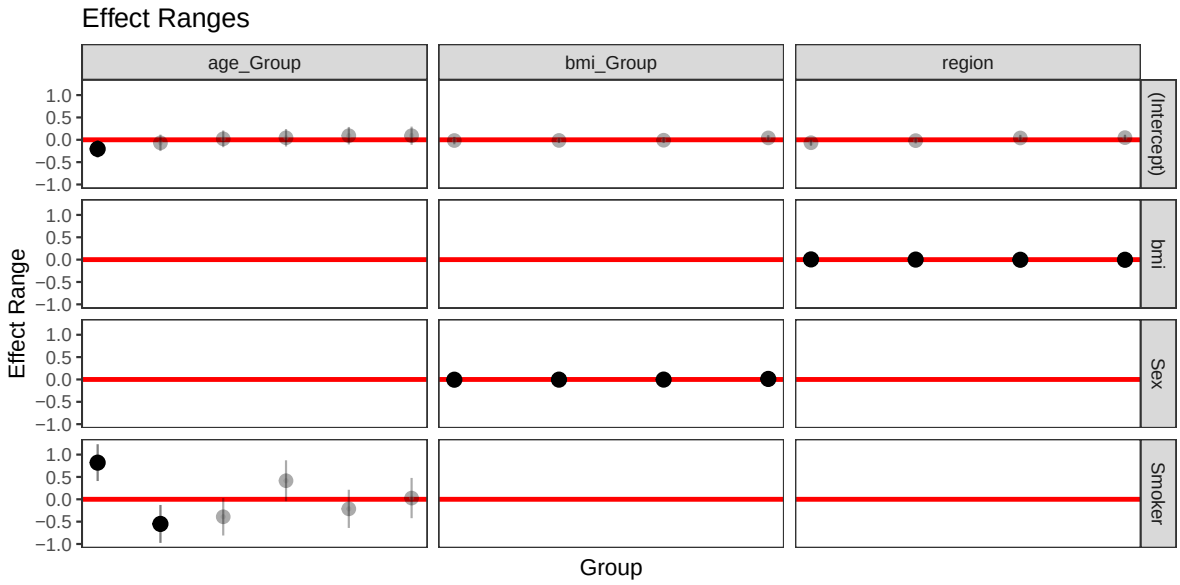


Figure 7: Visualization of random effects

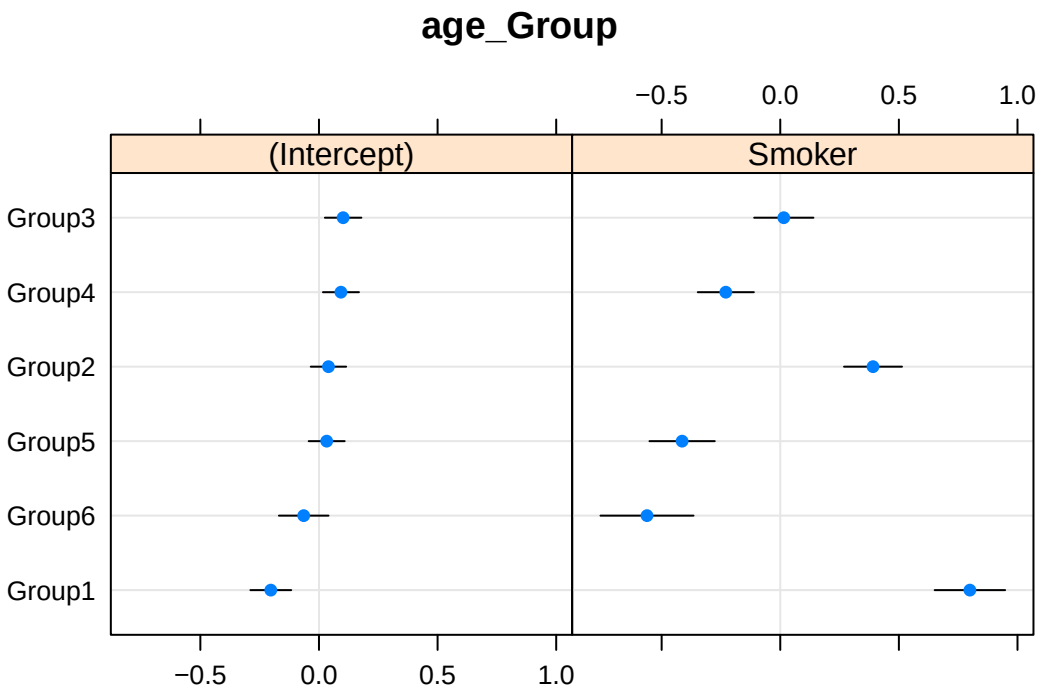


Figure 8: Visualization of random effects

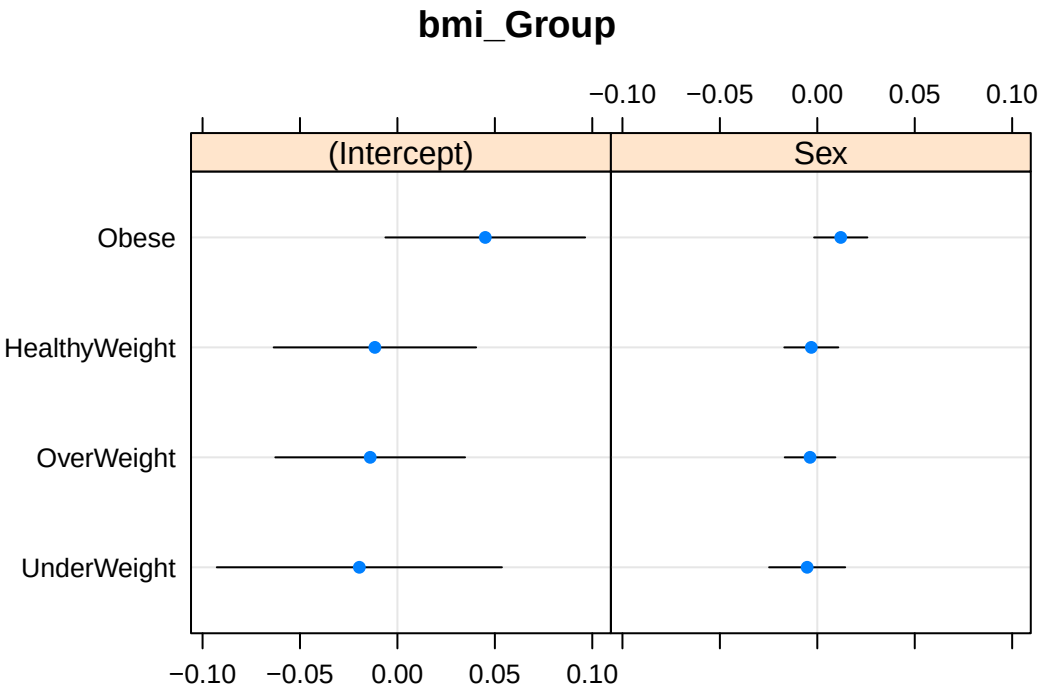


Figure 9: Visualization of random effects

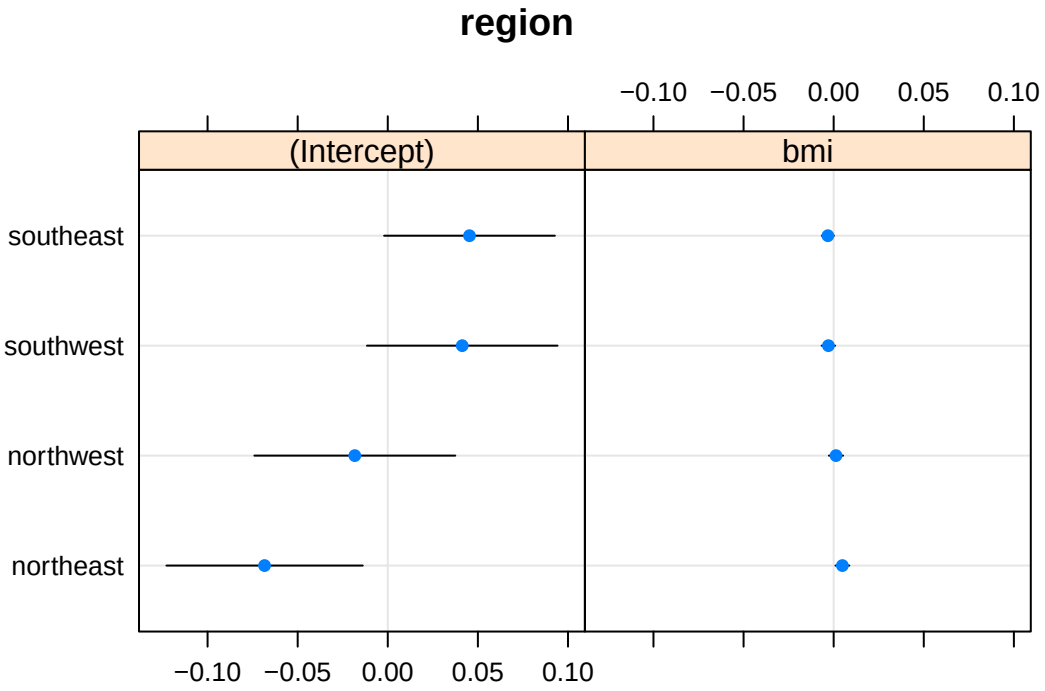


Figure 10: Visualization of random effects