**ST 511 Final Project Report**

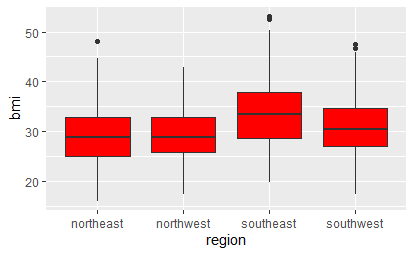
**Haotian Jia**

**Winter 2021**

***I. Introduction***

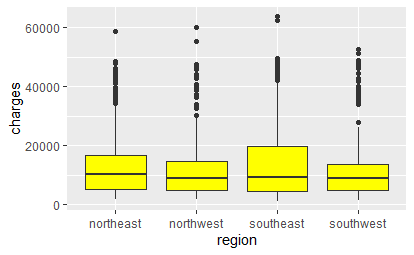
In this final project, I got the data set from the **“Kaggle.com”**, and the title of the data set is ***"Medical Cost Personal Datasets"***. Therefore, the proposal of this experiment is to test whether **BMI** (Body mass index, ideally 18.5 to 24.9) and **Insurance charges** are related to different **Regions** (Northeast; Northwest; Southeast; Southwest).

Figure 1 shown below is the box plot of BMI and Regions.



***Figure 1: The relationship between BMI and Regions***

Figure 2 shown below is the box plot of Charges & Regions.



***Figure 2: The relationship between Insurance charges and Regions***

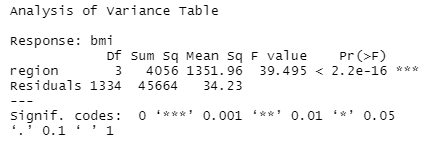
***II. Methods***

Since I want to test whether “BMI (Body mass index)” and “Insurance Charges” are related to different regions, so I prefer to use **“ANOVA”** to test the data because of multiple samples. Therefore, I made the following settings:

1. The null hypothesis is ;
2. The alternative hypothesis is ;

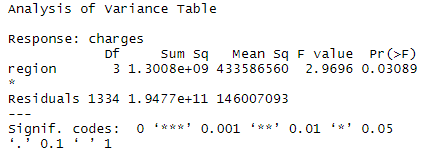
***III. Results***

**⑴ BMI:** Based on the data of Figure 3 below, we can get that that the “p-value” is . Besides, since , then we can know that the “p-value” is smaller than . Therefore, I will **reject the null hypothesis**, which means that there are at least two population means are different.



***Figure 3: ANOVA of “bmi ~ region”***

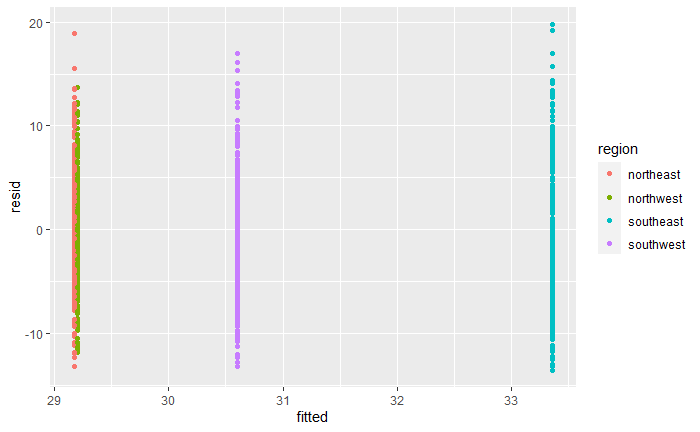
**⑵ Insurance Charges:** In addition, based on the data of Figure 4, I found that the “p-value” is , which is smaller than . Therefore, I will **reject the null hypothesis**, which means that there are at least two population means are different.



***Figure 4: ANOVA of “charges ~ region”***

***IV. Assessment***

Figure 5 and Figure 6 are showing the residual plots.



***Figure 5: The residual plot (BMI)***

***图表

描述已自动生成***

***Figure 6: The residual plot (Insurance Charges)***

***V. Conclusion***

Based on the analysis of the above four parts, I think the Null hypothesis which I set up is problematic, and it can be intuitively seen that there are many outliers in Figure 2. Therefore, I think I might consider deleting records with outliers in future areas for study or improving this final project by other means.