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.

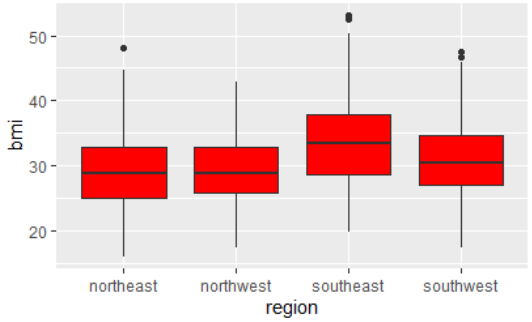

Figure 1: The relationship between BMI and Regions

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

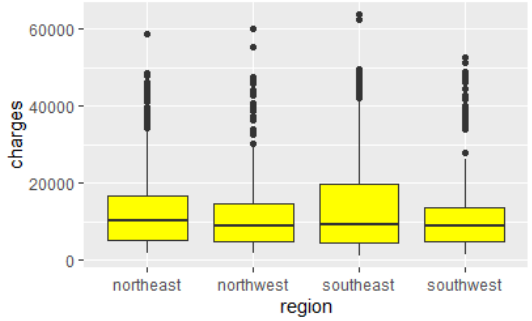

Figure 2: The relationship between Insurance charges and Regions

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 $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$;
2. The alternative hypothesis is $H_A: at\ least\ two\ population\ means\ are\ different$;
3. $\alpha = 0.05$

III. Results

(1) **BMI:** Based on the data of Figure 3 below, we can get that that the "p-value" is $2.2 * 10^{-16}$. Besides, since $\alpha = 0.05$, then we can know that the "p-value" is smaller than $\alpha = 0.05$. Therefore, I will **reject the null hypothesis**, which means that there are at least two population means are different.

```
Analysis of Variance Table

Response: bmi
             Df Sum Sq Mean Sq F value    Pr(>F)
region        3   4056 1351.96  39.495 < 2.2e-16 ***
Residuals 1334  45664   34.23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

Figure 3: ANOVA of "bmi ~ region"

(2) **Insurance Charges:** In addition, based on the data of Figure 4, I found that the "p-value"

is 0.03089, which is smaller than $\alpha = 0.05$. Therefore, I will **reject the null hypothesis**, which means that there are at least two population means are different.

```
Analysis of Variance Table

Response: charges
            Df     Sum Sq   Mean Sq F value  Pr(>F)
region       3 1.3008e+09 433586560  2.9696 0.03089
*
Residuals 1334 1.9477e+11 146007093
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
```

Figure 4: ANOVA of “charges ~ region”

IV. Assessment

Figure 5 and Figure 6 are showing the residual plots.

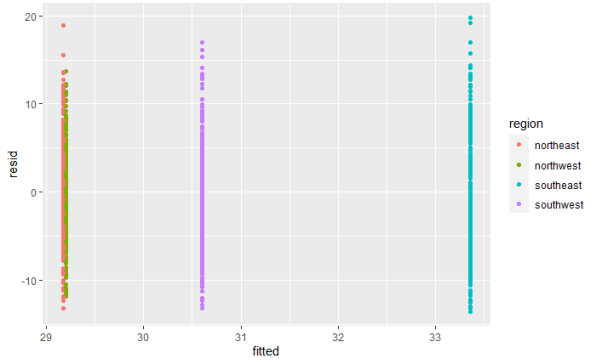

Figure 5: The residual plot (BMI)

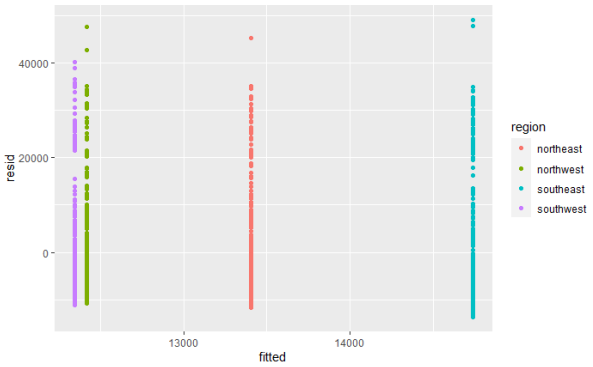

Figure 6: The residual plot (Insurance Charges)

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.