Analysis of Variance on the Effect of Smoking Status and Number of Children on US Health Insurance

San Jose State University

Le Dao (le.h.dao@sjsu.edu)
Dang Minh Nhu Nguyen (dang.m.nguyen@sjsu.edu)
Nhung Luong (nhung.luong@sjsu.edu)

MATH 161B: Applied Probability and Statistics II

Dr.Issa Tahir Bachar

May 14, 2022

Table of Contents

Abstract
Health care services are influenced by many factors, including demographic, social, economic, and health status [3]. Thus, the price charged for health care varies depending on many aspects such as age, sex, and so on. This study aimed to get a better understanding of underwriting risk by checking certain factors' influence on health insurance charges in the United States. Analysis was performed by using R programming language in RStudio with two-way ANOVA method on two factors: "children" and "smoker" in an open-sourced dataset "US Health Insurance" from Kaggle. Data cleaning and data transformation were performed to ensure the ANOVA model passes all assumptions, thus, we could further conduct Tukey's test to interpret the results. From the results, we concluded that there was no significant effect of "children" on the average insurance price. However, there was a significant impact of "smoker" on the average price being charged for health insurance in the United States. In addition, no relationship between "children" and "smoker" was found from the study. For further improvement, MANOVA will be implemented to utilize all data in the dataset so we can have a better conclusion on the stated problem.

I. Introduction

Analysis of variance or ANOVA analysis is a useful and important experimental method. It is widely used in research - quantitative analysis especially in research fields including biology, economics and psychology. There are different types of ANOVA including: one way, two way and three way. In case of one-way ANOVA, scientists perform an F-test on an independent factor. Meanwhile in two-way ANOVA, multiple tests can be developed in order to not only test the effects of any single factor to the hypothesis, but also an interaction among them.[5].

Decision making is always a focus topic to all real life aspects, and humans follow a variety of methods to identify its risk as well as trying to understand the effect of those variables. In fact, choosing the right health insurance is also a difficult decision for many people due to many factors that affect the price. It turns out there are underwriting risks that many people may not know. “Underwriting risk is the risk of uncontrollable factors or an inaccurate assessment of risks when writing an insurance policy” [8]. With the purpose of understanding the risk underwriting in Insurance Business, we had chosen a dataset involving US Health Insurance. [9]. Being able to clarify the goal and understand the importance of two-way ANOVA in testing the effect of the factors to decision making; in this project we applied two way ANOVA using R programming language to see if there was an effect of “children”, “smoker”, and the interaction of both factors on insurance price based on the dataset ‘US Health Insurance’ from Kaggle.

Below are the hypotheses that would be applied to three tests.

$$H_{0A}: \alpha_1 = \cdots = \alpha_I = 0 \quad \text{versus} \quad H_{aA}: \text{at least one } \alpha_i \neq 0$$
$$H_{0B}: \beta_1 = \cdots = \beta_J = 0 \quad \text{versus} \quad H_{aB}: \text{at least one } \beta_j \neq 0$$
$$H_{0AB}: \gamma_{ij} = 0 \text{ for all } i, j \quad \text{versus} \quad H_{aAB}: \text{at least one } \gamma_{ij} \neq 0$$

The result of our ANOVA should show whether or not the P value(s) are significant for each test, which means we should be able to reject the null hypotheses or not. Since many people buy insurance for family members, which means there are more risks in those plans, we expected the number of children would make a significant impact on insurance price. In terms of tobacco use, it was expected that if a person is a smoker, the person’s health plan is more expensive because the more likely he/she might need coverage on hospital or medicines. Lastly, we expect the effect of smoking on the insurance price and further to confirm the significant differences in price among smoker factors.

II. Materials and Methodology

1. Materials:

The dataset “insurance.csv” was downloaded from Kaggle, called ‘US Health Insurance Dataset’, depicts the amount of insurance being charged in the United States, including the significant factors for risk underwriting. The data has 1338 observations, where the “charge” for each insured individual is given against “Age”, “Sex”, “BMI”, “Children”, “Smoker”, and “Region”. Each of the features above is a mixture of numeric and categorical data. None features had been observed with missing variables.

In terms of programming language, nowadays, thanks to its ease-of-use and numerous useful libraries and packages, Python is the most commonly used programming language in data science and machine-learning projects. However, “Python is far behind the R programming language when it comes to general statistics and for this reason many scientists still rely heavily on R to perform their statistical analyses” [10]. Since this project is mainly focused on statistics, we chose R to have more tools specified for analyses. The programming language being implemented to assist in broadening the understanding about the relationship between “charges” and other variables in the dataset is R. Such packages like “data.table”, “dplyr”, “ggpubr” and “AID” were used in our R-code project.

The project was concentrated on the usage of two-way ANOVA. This method was implemented to evaluate if the difference between the means of three or more independent groups divided by two different features, or also known as factor, is statistically significant and whether there is an interaction between these two factors against the given response factor [12]. An interaction between factors means that the impacts of one or more specific factors have a multiplicative influence on other variables, as seen by your outcome variables. The two-way ANOVA is more complicated than the one-way ANOVA, but it still follows the same rules and needs to meet a few assumptions: normality, equal variance, and independence [12]. In insurance.csv, “children” was chosen to be factor A, and “smoker” was chosen to be factor B to predict their relationship with the insured charges.

Two-way ANOVA with replication required the dataset to meet a few assumptions: Normality, Equal Variances, Independence [12]. Next is the discussion on methodology used in this project.

2. Methodology:

a, Hypotheses:

In this project, we conducted a two-way ANOVA method on two factors: children and smokers with the following hypotheses. All codes are written in R using RStudio platform.

Our null hypotheses are:

- H_{OA} stated that the different levels of factor A (children) has no effect on the true average insurance price.
- H_{OB} stated that the different levels of factor B (smoker) has no effect on the true average insurance price.
- H_{OAB} stated that there is no interaction between the number of children at level i and factor smoker status at level j.

And corresponding alternative hypotheses are:

- H_{aA} stated that there is at least one level of factor A (children) that has an effect on the true average insurance price.
- H_{aB} stated that there is at least one level of factor B (smoker) that has an effect on the true average insurance price.
- H_{aAB} stated that there is interaction between the number of children at level i and factor smoker status at level j.

b, Data Cleaning:

Packages “data.table”, “dplyr”, “ggpubr”, “car” were needed for this project. First, import the file “insurance.csv” into the R console. After the data was loaded, it was crucial to check for the existence of missing values. Since the data didn’t have any NAs, no observation was omitted. The next step was to generate the table which contained the mean and standard deviation at each level of variables children and smokers. *Table 1* was the output of R.

To make it easier to comprehend the data in the table above, a box plot and a line plot were then generated from the package “ggpubr”. From the boxplot *Figure 1* and the standard deviation values in the table, the variance of each group varied greatly and it appeared that there was a possibility that the homoscedasticity assumption might be violated. To determine whether the assumption was violated or not, Levene test would be performed later in R.

Based on *Figure 2*, graph of smoker versus non-smoker, from 0-3, the smoker line and non-smoker line seems to parallel. Since the lines were parallel, it signaled the fact that the interaction between the two factors smoker and children might not exist. As mentioned above, *Figure 1* appeared to signal a mild violation of the homogeneity assumption; however, ANOVA is fairly robust against this violation if each group had the same sample size [11]. Hence, before conducting any test, a for loop was generated to determine the minimum number of observations in each group, which was 9. Then, package “data.table” was used to randomly sample from each group and resulted in the final dataset called new_data where there were 9 observations at each level of factors. The result is shown in *Table 2*. Because assumptions check of normality on the model did not pass with the original data, we transformed the data by using logarithm.

The assumption of homogeneity is important for ANOVA testing and in regression models. In ANOVA, when homogeneity of variance is violated, there is a greater probability of falsely rejecting the null hypothesis [2]. This project uses the residuals versus fits plot to check the homogeneity of variances. In the plot below, there is quite no evident relationship between residuals and fitted values (the mean of each group), which is good. So, we can assume the homogeneity of variances. *Figure 3.* We also applied Levene’s test of Homogeneity of Variance [4] using R to determine if our model violate the assumptions, the result in *Figure 4* states that its p-value is significantly larger than the threshold value (i.e. 0.05), we then can confirm that our model pass the homogeneity assumption.

The normal probability plot of residuals is used to verify the assumption that the residuals are normally distributed. The normal probability plot of the residuals should approximately follow a straight line. *Figure 5.* Our plot seemed to be forming a straight line except some points were a little off. In order to attain the normality assumption, we decided to perform the Shapiro-Wilk normality test [7] based on our anova residuals. From the result *Figure 6*, since p-value was also larger than 0.05 threshold, we believed that no indication that normality is not violated on our model.

c, Analysis of Variance:

The next step would be performing two-way ANOVA using R, the pseudocode is aov(formula, data = NULL, projections = FALSE, qr = TRUE,contrasts = NULL, ...)[1], in which “formula” will be in the form of the response variable and followed by 3 predictors (i.e. smoker, children,

and the interaction between smoker and children). The data will be the new one after the binding step, everything else should be left as default.

d, Result:

After running the two-way ANOVA test, we then check the summary, the detailed result should be found in *Figure 7.* From the result, with the default threshold set to 0.05, only the p-value of smoker is much smaller than alpha. Thus, we can conclude that:

- Fail to reject H_{oA}, the data suggests that there is no significant effect of "children" on insurance prices.
- Reject H_{oB}, the data suggests that there is a significant effect of "smoker" on insurance price.
- Fail to reject H_{oAB} , the data suggests that there is no significant effect of the interaction between "smoker" and "children" on insurance price.

e, Tukey's Test:

Identifying the impact of "smoker" on the insurance's price from the above ANOVA step, we then process another step to check if there is a significant difference in price within the group, to state it's different, between the people who smoke versus non smokers. Tukey's test detailed result, generated by R, is on *Figure 8.* Within the 95% confidence interval, the p-adjust output is very small which statistically indicates that there is a significant difference of insurance's price if the person's smoking or not.

IV. Conclusion

This study gave a better understanding of risk in Insurance business by checking two factors (children and smoker) influence over health insurance charges (charge) in the United States. Data cleaning, data transformation, and codes were performed in R to make sure our ANOVA passes assumptions of Normality, Equal Variances, and Independence.Based on the results from ANOVA and Tukey's test, firstly we can conclude that we failed to reject H_{oA}, which is not expected, meaning there was no significant effect of "children" on the average insurance price charged. Secondly, we rejected H_{oB} as expected, that means there was a significant impact of Smoker on the average insurance price in the United States. Lastly, we rejected H_{oAB} as expected to indicate that there was no relationship between the number of children and smoker status found from the study.

In this project, the most difficult part was finding the dataset that is suitable for using ANOVA due to the assumptions of normality and homogeneity of variance assumption. We had chosen many data sets, and they could not pass the normality test because the normal probability plot of the residuals did not approximately follow a straight line. In fact, most dataset available online for analysis are observational, and they are not normally distributed. Thus, many techniques such as dropping outliers and using logarithms have been used in order to have enough evidence for the normality assumptions. When all assumptions are reasonably passed, we can safely use ANOVA method for analysis, conduct Tukey's test based on significant variables, and show results as above. From that, our next step is to work on implementing MANOVA and hopefully can utilize all information given in the dataset. With that, we can give better judgment on the issue of underwriting risks in the insurance business.

V. Appendix

	smoker	children	mean	sd
1	no	0	9476.782	6232.324
2	no	1	6834.396	3697.116
3	no	2	10660.357	6538.371
4	no	3	10206.662	7401.914
5	yes	0	27195.365	12795.606
6	yes	1	28804.452	9997.936
7	yes	2	34339.513	10141.498
8	yes	3	26703.541	10768.286

Table 1: Mean and standard deviation of Smoker and Number of Children

Smoking status/Number of children	0	1	2	3
No	9	9	9	9
Yes	9	9	9	9

Table 2: Filtered data

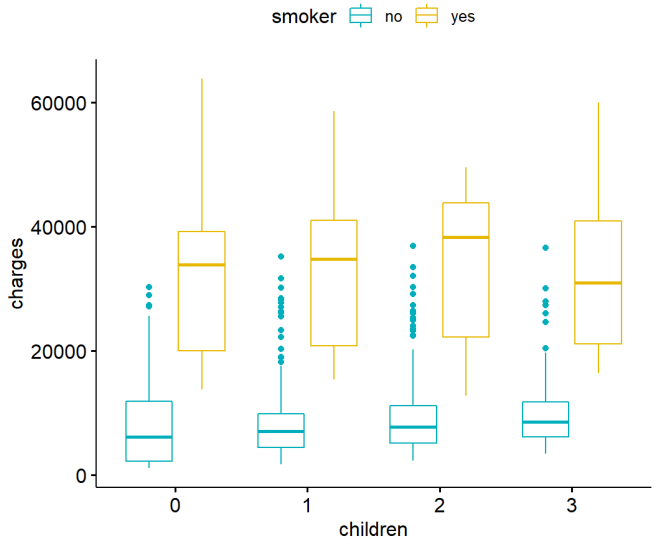

Figure 1: Box plot of children vs charges grouped by smoker

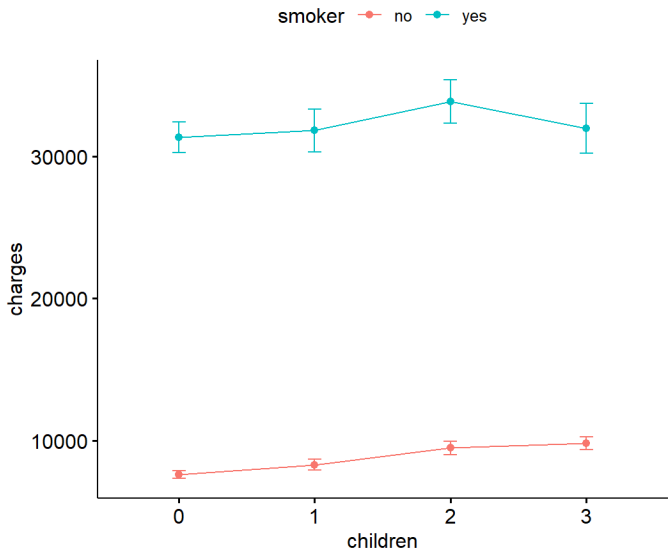

Figure 2: Line plot of children vs charges grouped by smoker

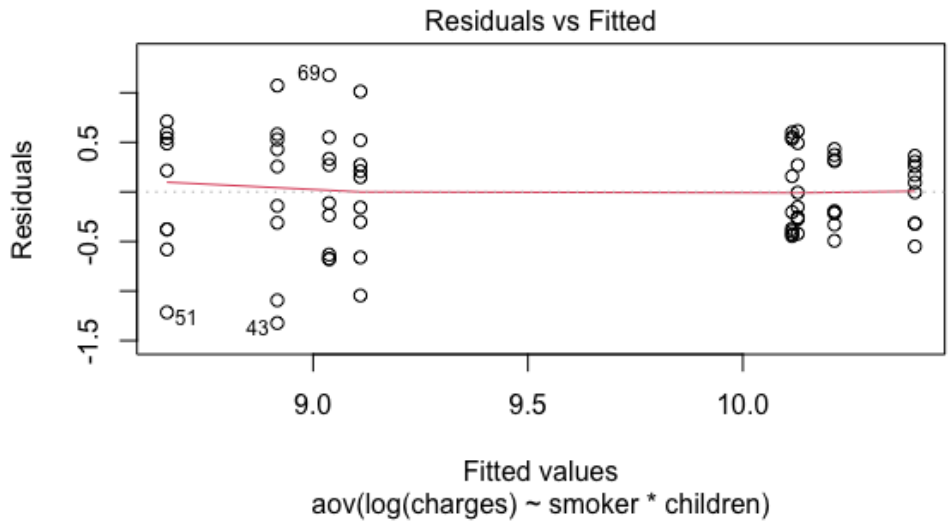

Figure 3: Residual versus Fitted values for our ANOVA model

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  7  1.3198 0.2556
```

Figure 4: Levene's Test result

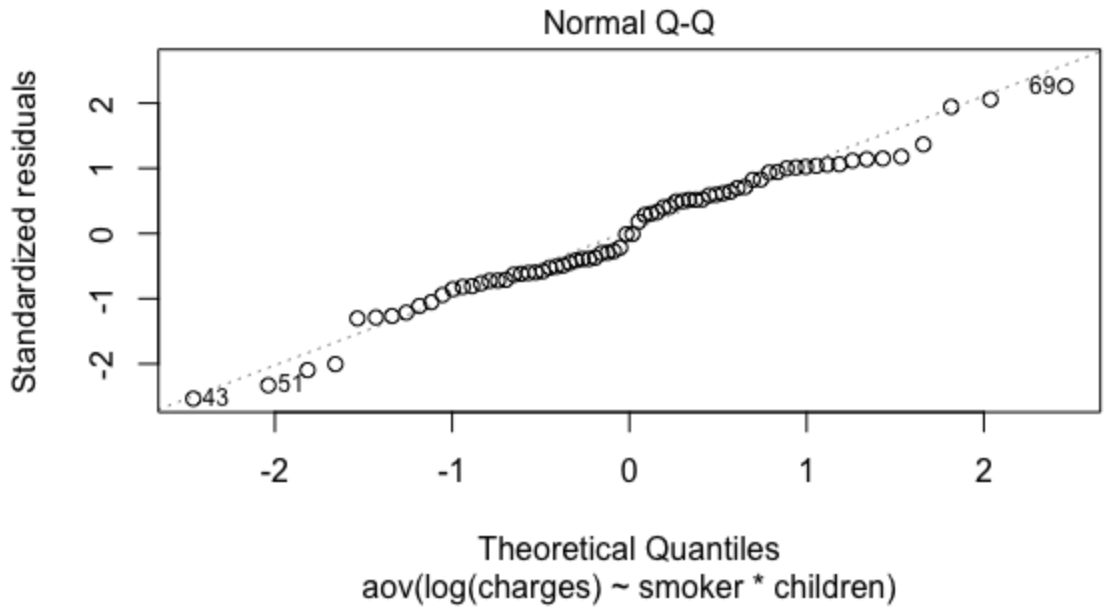

Figure 5: Normal QQ plot

```
        Shapiro-Wilk normality test

data:  aov_residuals
W = 0.97681, p-value = 0.2046
```

Figure 6 : Shapiro-Wilk test result

```
                Df Sum Sq Mean Sq F value   Pr(>F)
smoker           1 29.628  29.628  96.752 2.01e-14 ***
children         3  0.993   0.331   1.081    0.363
smoker:children  3  0.529   0.176   0.575    0.633
Residuals       64 19.598   0.306
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: ANOVA Summary of Test Result

```
             diff      lwr     upr        p adj
yes-no 1.282963 1.022396 1.54353 9.577783e-12
```

Figure 8: Tukey’s Test Result

VI. References

[1] *aov function - RDocumentation*. (2021). R Documentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aov

[2] *Assessing the Assumptions of Homogeneity · UC Business Analytics R Programming Guide*. (2017). University of Cincinnati. https://uc-r.github.io/assumptions_homogeneity#:%7E:text=The%20assumption%20of%20homogeneity%20is,to%20residuals%20(aka%20errors)

[3] Kong, N. Y. (2020, July 20). *Factors influencing health care use by health insurance subscribers and medical aid beneficiaries: a study based on data from the Korea welfare panel study database - BMC Public Health*. BioMed Central. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-09073-x

[4] *levene.test function - RDocumentation*. (2020). R Documentation. https://www.rdocumentation.org/packages/lawstat/versions/3.4/topics/levene.test

[5] Rouder, J. N. (2016, April 11). *Model comparison in ANOVA*. SpringerLink. https://link.springer.com/article/10.3758/s13423-016-1026-5?error=cookies_not_supported&code=b76e3032-297f-490d-b71f-df1f7d1745d3

[6] *RPubs - 2-Way ANOVAs in R*. (2017, November 4). Rpubs. https://rpubs.com/tmcurley/twowayanova

[7] *shapiro.test function - RDocumentation*. (2021). R Documentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/shapiro.test

[8]*Underwriting Risk*. (2021, May 29). Investopedia. https://www.investopedia.com/terms/u/underwriting-risk.asp#:%7E:text=Key%20Takeaways,than%20it%20receives%20in%20premiums

[9] *US Health Insurance Dataset*. (2020, February 16). Kaggle. https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset

[10] Vallat, R. (2018, November 19). *Pingouin: statistics in Python*. Journal of Open Source Software. https://joss.theoj.org/papers/10.21105/joss.01026

[11] Z. (2021a, April 11). *What is the Assumption of Equal Variance in Statistics?* Statology. https://www.statology.org/equal-variance-assumption/

[12] Z. (2021, November 30). *Two-Way ANOVA: Definition, Formula, and Example*. Statology. https://www.statology.org/two-way-anova/