VIETNAM NATIONAL UNIVERSITY - HCM
Ho Chi Minh City University of Technology
Faculty of Computer Science and Engineering

PROBABILITY AND STATISTICS (MT2001)

Project Report:

Data Analysis

Instructor: Mr. Nguyễn Tiến Dũng
Students: Ôn Quân An - 1852221
Lê Thiên Ân - 1852255
Trần Quốc Anh - 1852247
Đào Đức Bảo - 1852258

Ho Chi Minh City, June 2020

Contents

1 Dataset Description

Medical Cost Personal Datasets

- This Data is practically is used in the book "Machine Learning with R by Brett Lantz"; which is a book that provides an introduction to machine learning using R. (https://www.amazon.com/Machine-Learning-R-Brett-Lantz/dp/1782162143).
- All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book. The following data obtained from Kaggle, explain the cost of a small sample of US population Medical Insurance Cost based on some attributes depicted below.

1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622
7	46	female	33.440	1	no	southeast	8240.590
8	37	female	27.740	3	no	northwest	7281.506
9	37	male	29.830	2	no	northeast	6406.411
10	60	female	25.840	0	no	northwest	28923.137
11	25	male	26.220	0	no	northeast	2721.321
12	62	female	26.290	0	yes	southeast	27808.725
13	23	male	34.400	0	no	southwest	1826.843

(1325 rows remaining)

Content of dataset:

- **age**: age of primary beneficiary
- **sex**: insurance contractor gender, [female, male]
- **bmi**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m^2) using the ratio of height to weight.
- **children**: Number of children covered by health insurance / Number of dependents
- **smoker**: Smoking, [yes, no]
- **region**: the beneficiary's residential area in the US, [northeast, southeast, southwest, northwest]
- **charges**: Individual medical costs billed by health insurance

Type of attributes:

- "**age**" is ordinal, "**sex**" is categorical, "**bmi**" is quantitative, "**children**" is quantitative, "**smoker**" is categorical, "**region**" is categorical, "**charges**" is quantitative.

2 Statistical Method

2.1 Linear Regression

2.1.1 Definition

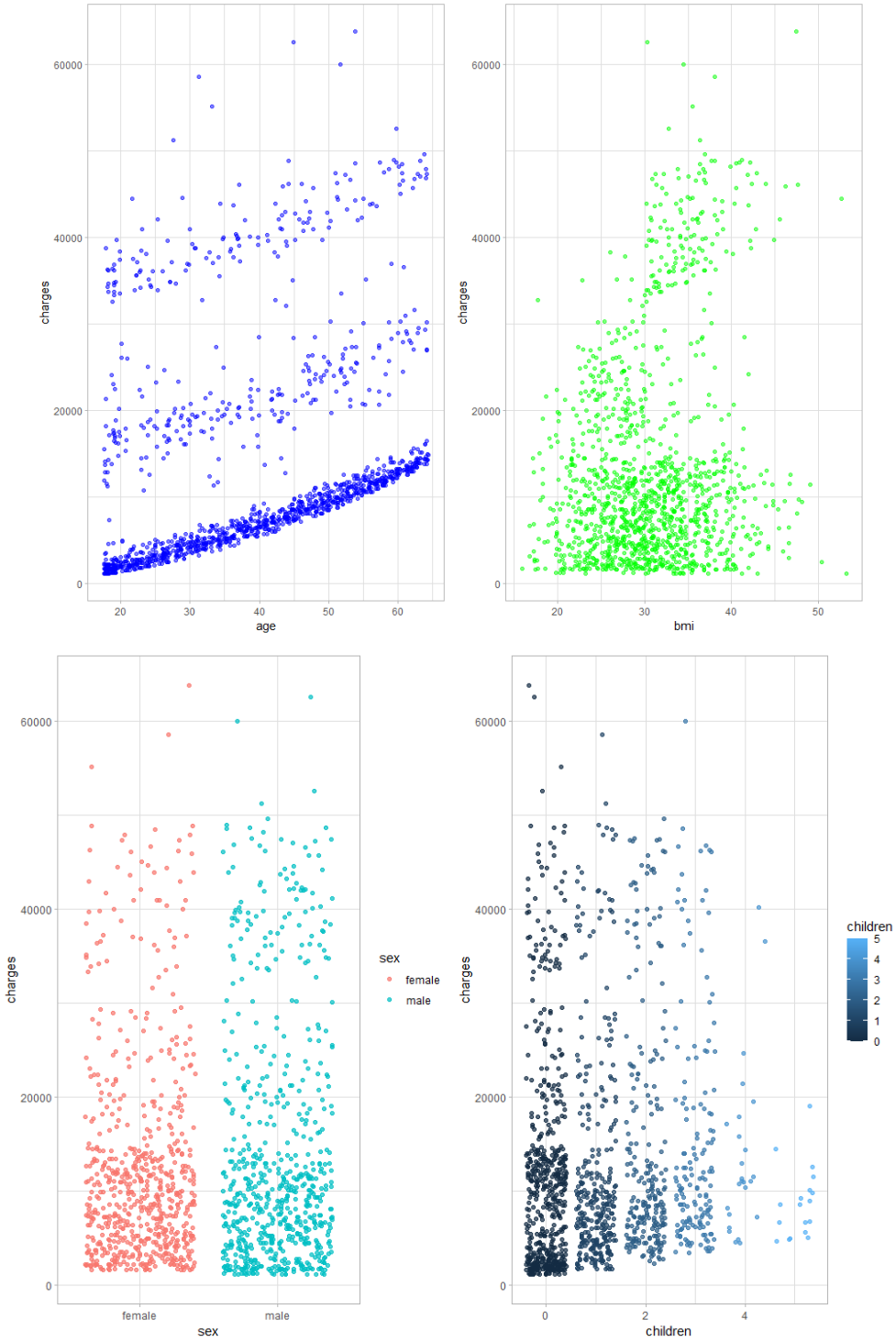
In statistics, **linear regression** is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called **multiple linear regression**. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

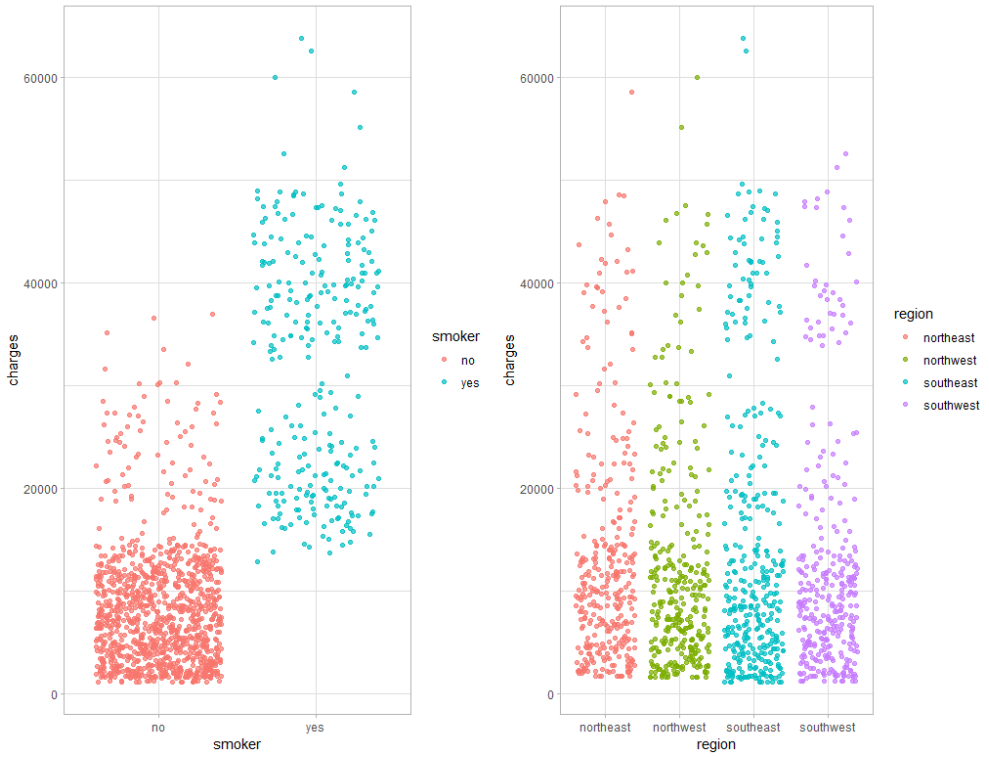
2.1.2 Exploratory Data Analysis

The dataset depicts information about a person which is categorized into seven categories including **age**, **sex**, **BMI**, **children**, **smoker**, **region** and **charges**. After going through the dataset for some times, we concluded that it is rather reasonable to assume that the variable charges is influenced by the rest six variables. Therefore, we are going to check if the dataset is a suitable candidate for the Linear Regression data analysis method and if our assumption is satisfiable.

Initially, we need to check the correlation between our targeted variable (**charges**) and the other independent variables (**age**, **sex**, **BMI**, **children**, **smoker**, **region**).

Correlation between Charges and Age/BMI/Children/Smoker/Region





Summary:

1. As Age and BMI go up Charges for health insurance also trends up.
2. No obvious connection between Charges and Sex. Charges for insurance with 4-5 children covered seems to go down.
3. Charges for Smokers are higher than non-smokers. No obvious connection between Charges and Region.

2.1.3 Implementation

Hypothesis representation:

We will use x_i to denote the independent variable and y_i to denote dependent variable. For Simple Linear regression, a pair of (x_i,y_i) is called training example. The subscript i in the notation is simply index into the training set. We have m training example then i=1,2,3,...m.
The hypothesis function represented as:

$$h_\theta(x_i) = \theta_0 + \theta_1 * x_i$$

- θ_0, θ_1 are parameters of hypothesis

However, as our assumption suggests, the target variables will presumably be influenced by the other six variables. Hence, applying Simpler Linear Regression on this dataset will not be satisfiable for the data analysis process as well as forecasting the possible trend.

To tackle this, we will be adopting a more advanced and reliable method, namely, Multiple Linear Regression method. In this method, we will denote the independent variables by x_i and y_i will be the dependent variable. We have n independent variable then j = 1,2,3...n.

$$h_\theta(x_i) = \theta_0 + \theta_1{*}x_{i1} + \theta_2{*}x_{i2} + \ldots + \theta_j{*}x_{ij} + \ldots + \theta_n{*}x_{mn}$$

- θ_0, θ_1, ... , $\theta_j \ldots \theta_n$ are parameter of hypothesis.
- m Number of training examples.
- n Number of independent variables.
- x_{ij} is i^{th} training example of j^{th} feature.

Now for the real work. To judge how how good our model is, we need something to test it against. We can accomplish this using a technique called cross-validation. Cross-validation can get much more complicated and powerful, but in this example, due to being inexperienced as well as to avoid undesirable errors, we are going do the most simple version of this technique.
Steps:

1. Divide the dataset into two datasets: A 'training' dataset that we will use to train our model and a 'test' dataset that we will use to judge the accuracy of that model.
2. Train the model on the 'training' data.
3. Apply that model to the test data's X variable, creating the model's guesses for the test data's Ys.
4. Compare how close the model's predictions for the test data's Ys were to the actual test data Ys.

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. With these step, we can avoid problems like overfitting.

Splitting the Data

Here, we divide the dataset into two parts, 70% random data for training and 30% left for testing.

```
split <- sample.split(data, SplitRatio = 0.7)
Data_train <- subset(data, split = "TRUE")
Data_test <- subset(data, split = "FALSE")
```

Train the Model

There are seven independent variables. The target variable here is **charges** and remaining six variables: **age**, **sex**, **bmi**, **children**, **smoker**, **region** are independent variable. There are multiple independent variables, so we need to fit Multiple linear regression. Then the hypothesis function should looks like:

$$h_\theta(x_i) = \theta_0 + \theta_1{*}\text{age} + \theta_2{*}\text{sex} + \theta_3{*}\text{bmi} + \theta_4{*}\text{children} + \theta_5{*}\text{smoker} + \theta_6{*}\text{region}$$

In order to forecast the possible values for the dependent variable, we will be using our model parameter for the train dataset. Thereafter, comparisons between the model and the actual value in test set will be made. We compute **Mean Square Error** using formula:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

R^2 is statistical measure of how close data are to the fitted regression line. R^2 is always between 0 to 1. 0 indicated that model explains none of the variability of the response data around it's mean. 1 indicated that model explains all the variability of the response data around the mean.

$$R^2 = 1 - \frac{SSE}{SST}$$

$$\text{SSE (Sum of Square Error)} = \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

$$\text{SST (Sum of Square Total)} = \sum_{i=1}^{m} (y_i - \bar{y}_i)^2$$

But RStudio has provided us with an excellent and simple-to-use function to calculate these things:

```
formula_0 <- as.formula("charges ~ age + sex + bmi + children + smoker + region)
model_0 <- lm(formula_0, data = Data_train)
```

We then get the summary of the created model.

```
Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value   Pr(>|t|)
(Intercept)      -11938.5      987.8  -12.086   < 2e-16 ***
age                 256.9       11.9   21.587   < 2e-16 ***
sexmale            -131.3      332.9  -0.394   0.693348
bmi                 339.2       28.6  11.860    < 2e-16 ***
children            475.5      137.8   3.451   0.000577 ***
smokeryes         23848.5      413.1  57.875    < 2e-16 ***
regionnorthwest    -353.0      476.3  -0.741   0.458769
regionsoutheast   -1035.0      478.7  -2.162   0.030782 *
regionsouthwest    -960.0      477.9  -2.009   0.044765 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,     Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

According to the summary, it is evident that smoking seems to have a great influence on the **charges** whereas **sex** does not leave much impact on the cost. From this information, we can conclude that the dataset is a potential candidate for Linear Regression. Hence, to simplify the problem, we will then create another model without the non-significant variables and check if the performance can be improved. Ultimately, we have the new simplified model which has only **age**, **bmi**, **children**, **region** and **smoker**.

```
formula_1 <- as.formula("charges ~ age + bmi + children + smoker + region)
model_1 <- lm(formula_1, data = Data_train)
```

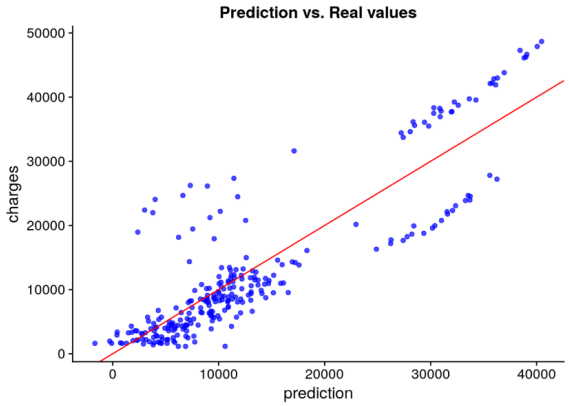
We then compare the new model with the original one.

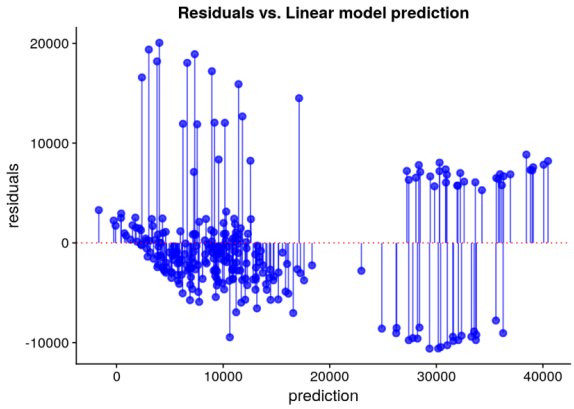
- Adjusted R-squared for first model: 0.7494
- Adjusted R-squared for second model: 0.7496
- RMSE for first model: 6041.68
- RMSE for second model: 6042.03

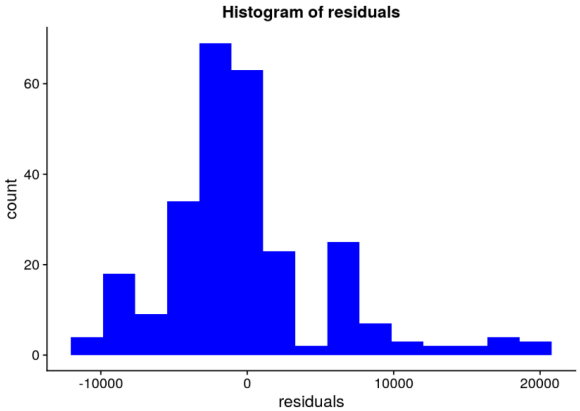
As we can see, the performance of the second model does not differ much from the original model. Hence, we will use the second model since it is simpler.

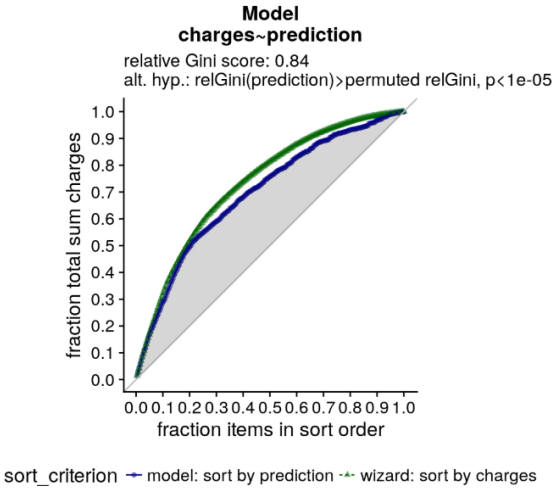
2.1.4 Model Performance

Now that we have trained the model, to determine the predicting performance of the trained model, comparisons with the test data will be made.









As the graphs depict, the model predicts fairly well since most of the residuals allocate close to 0. However, as the prediction values increases, the residuals also varies slightly. But overall, the model still performed acceptably well.

Finally, to confirm the model's performance, some examples will be given to test the model. Three different people with different information will be given to see how much the Insurance charges on health care will be for each of them.

Person 1: Nguyen Van A, 20 years old, BMI 25.4, no children, smokes, from northwest. Health care charges for Nguyen Van A: 25235.4
Person 2: Tran Duc B, 56 years old, BMI 49, 3 children, doesn't smoke, from northeast. Health care charges for Tran Duc B: 19904.57
Person 3: Dang Van C, 30 years old, BMI 31.2, 1 children, smokes, from southeast. Health care charges for Dang Van C: 29561.78

2.1.5 The Result

From the performance of our testing model and the results of the prediction shown above, "sex" attribute does not affect much on insurance charges, but being a smoker influences heavily on the amount of charge one person will be charged, which is satisfiable to our initial assumption.

2.2 Two-way Analysis of Variance (ANOVA)

2.2.1 Definition

Analysis of variance (ANOVA) is a statistical method that separates observed aggregate variability found inside a dataset into two basic components: dependent factors and independent factors. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study. The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

2.2.2 Implementation

Hypothesis Representation In this statistical method,there are three null hypotheses we need to test:

- There is no difference in the means of observations grouped by one factor
- There is no difference in the means of observations grouped by the other factor
- There is no interaction between two factors

Data preparation In order to conduct this method in R, we sample a group of 460 rows which have full attributes.In this study, we want to evaluate if there is a significant two-way interaction between sex and smoker factor on explaining the total charges. An interaction effect occurs when the effect of one independent variable on an outcome variable depends on the level of the other independent variables.

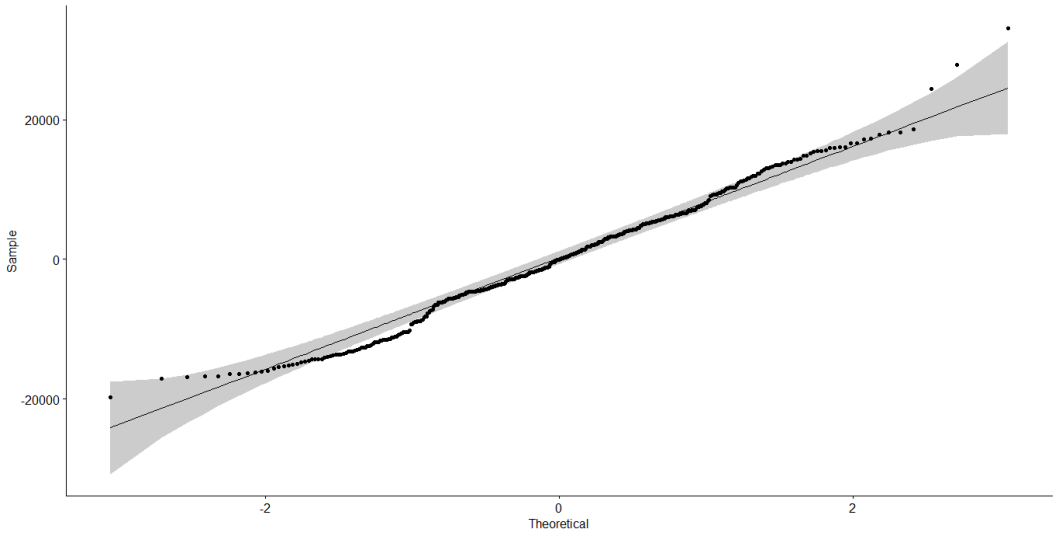
2.2.3 Checking the validity of the test

Since ANOVA assumes that the data are normally distributed and the variance across groups are homogeneous. We can check that with some diagnostic plots.

Normality plot of the residuals. In the plot below, the quantiles of the residuals are plotted against the quantiles of the normal distribution. A 45-degree reference line is also plotted.

The normal probability plot of residuals is used to verify the assumption that the residuals are normally distributed.

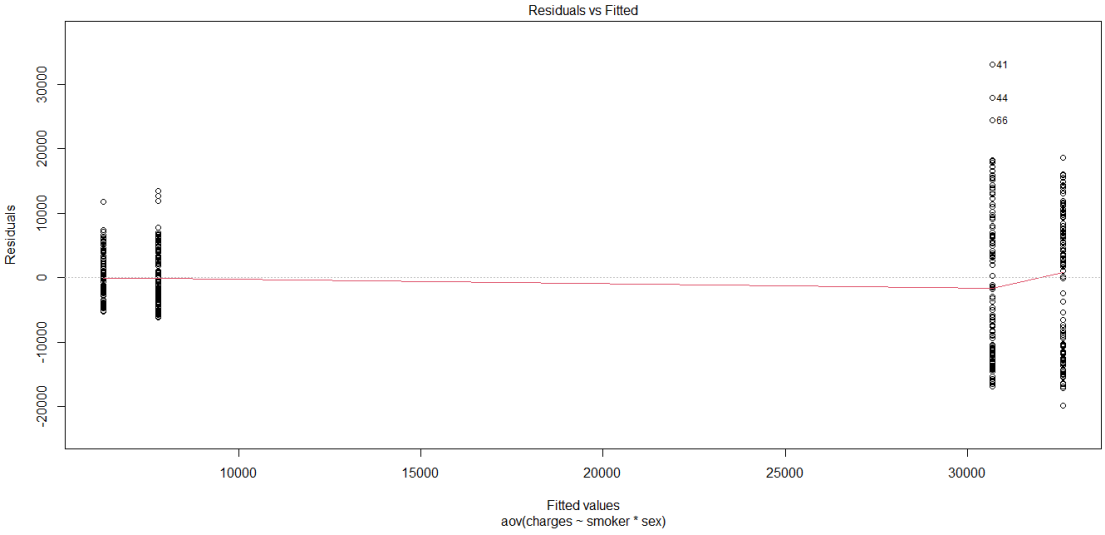
The normal probability plot of the residuals should approximately follow a straight line.



As all the point follow a straight line, we can assume normality for our dataset. We do not conduct Shapiro-Wilk test(which is very popular when testing normality) because the sample size in this case is very big since for large amounts of data even very small deviations from normality can be detected, leading to rejection of the null hypothesis event though for practical purposes the data is more than normal enough.

Homogeneity check

The **residuals versus fits plot** is used to check the homogeneity of variances. In the plot below, there is no evident relationships between residuals and fitted values (the mean of each groups), which is very good. Linear property seems to hold reasonably well as the red line is close to the dashed line. Therefore we can assume the homogeneity of variances.



2.2.4 Computation

To comprehend the analysis of variance model, we create a model named modelaov by using function aov() in R studio. And then we fit the model in function Anova() with type "III" sums of squares. After computing, the program displays a table of information as above figure, each column with different value.

```
Df    Sum Sq   Mean Sq F value Pr(>F)
sex           1 2.743e+07 2.743e+07   0.367 0.5447
smoker        1 6.724e+10 6.724e+10 900.865 <2e-16 ***
sex:smoker    1 3.290e+08 3.290e+08   4.407 0.0364 *
Residuals   442 3.299e+10 7.464e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Sum Sq: These are the sum of squares for all rows.
- Df: This is the degrees of freedom.
- F value: This is the F statistic associated with the given source.
- Pr(>F): This is the p-value associated with the F statistic of a given source.

2.2.5 The Result

A two-way ANOVA was conducted to examine the effects of sex and smoker on insurance charges.

Residual analysis was performed to test for the assumptions of the two-way ANOVA.

There were no extreme outliers, residuals were normally distributed and there was homogeneity of variances.

From ANOVA table, we can conclude there was statistically significant interaction between sex and smoker on total insurance charges (p-value $= 0.0364 < 0.05$) which indicate that the relationships between "sex" and insurance charges depends on the smoker factor . "smoker" factor effect is statistically significant (p-value $< 2{*}10^{-16}$) while "sex"'s effect is very weak(p-value =0.5547). The result proves that whether the person is smoker or not will impact significantly the mean insurance charges.

3 Discussion and Conclusion

Throughout the project, the main aim of our research is to understand the relationships between independent variables in a dataset by applying statistical data analysis methods. From the outcome of both methods (Linear Regression and ANOVA), it is evident that being a smoker, regardless of sex, will influence heavily on a person's health and, therefore, results in a drastic increase in the personal medical charges.

This study also found that there is a considerable increase in the cost of medical care as a person's BMI record goes up. Hence, it is advisable that one stay in shape and try to keep their BMI in the optimal and healthy degree to assure a healthy lifestyle as well as to maintain their budget low as low as possible.

One of the limitations of this study is that we selected R to perform the data analysis since, comparing to Python, R is much more complicated, which resulted in us taking more time than expected to do research about the language as well as browsing for suitable packages to graph the data suitably. Moreover, there are still much more to be improved in the way we access the data such as separating the data in to the smoker and non-smoker categories to do further analysis. Nevertheless, we still managed to graph the data in an easy-to-understand way and did the data analysis reasonably well.

Bringing up the rear, this study provided evidence that a person should stay in his/her healthy shape as well as say no to smoking as a hobby to experience a better life.

References

[1] *Linear Regression Tutorial*, Kaggle,
https://www.kaggle.com/sudhirnl7/linear-regression-tutorial

[2] *Linear regression*, Wikipedia,
https://en.wikipedia.org/wiki/Linear_regression

[3] *Health Care Cost Prediction w/ Linear Regression*, Kaggle,
https://www.kaggle.com/ruslankl/health-care-cost-prediction-w-linear-regression

[4] *Analysis of variance*, Wikipedia,
https://en.wikipedia.org/wiki/Analysis_of_variance

[5] *Two-way analysis of variance*, Wikipedia,
https://en.wikipedia.org/wiki/Two-way_analysis_of_variance

[6] *Two-Way ANOVA Test in R*, STHDA,
http://www.sthda.com/english/wiki/two-way-anova-test-in-r