**Insurance Premium Prediction / Linear Regression**

## 1. Research scenario and question(s)

Research Scenario -

This dataset contains insurance data for policyholders. Each row represents one policyholder and for each policyholder, we have age, gender, BMI, number of children, smoker/non-smoker, geographic region, and premium amount charged.

Our goal is to analyze the relationship between independent variables - age, gender, BMI, number of children, smoker/non-smoker, and the response variable – premium-charged.

We fit a multiple linear regression (MLR) model and perform regression diagnostics to analyze the fit. If the fit is reasonable, we perform the global F-test and subsequent t-test to determine which of the explanatory variables are significant predictors of the premium-amount charged.

Research Question(s) –

1. Is there a linear relationship between the explanatory variables and the response variable?
2. If yes, are age, gender, BMI, number of children, smoker/non-smoker significant predictors of the premium-amount-charged to a policyholder?
3. If yes, what is the relative contribution of these independent variables in predicting the premium-amount that will be charged to a policyholder?

## 2. Description of the data set

This dataset is a publicly available dataset and is obtained from this location on Kaggle - <https://www.kaggle.com/simranjain17/insurance>

The original dataset contains information about 1338 policyholders. For each policyholder, we have the following information –

* Age : The age of policyholders range from 18 – 64 years.
* Gender : Female/Male
* BMI : Body Mass Index for the policyholders ranges between 15.96 – 53.13
* Children : refers to the number of children the policyholder has – ranges between 0-5.
* Region : refers to the geographic location of the policyholder – northeast, northwest, southeast & southwest.
* Premium-charged : refers to the premium amount charged to the policyholder – ranges between $1121.8 – $63770.4

Note: I will not be using the “region” variable for my analysis.

Sampling : I have taken a stratified sample of 1000 datapoints containing 250 records for each of the 4 regions.

Data Cleaning : The dataset does not contain any missing values or noisy data.

## 3. Statistical methods used

Multiple Linear Regression(MLR) to model the nature of linear relationship between age, bmi, gender, number-of-children, and response variable – premium-charged.

I will then perform regression diagnostics to determine if the MLR model was a good fit for the data.

If the assumptions of linear regression are met, I will perform the global F-test to determine if these independent variables are significant predictors of premium-charged.

I will then perform t-test to determine the relative contribution of each of these independent variables in predicting the premium-charged.

## 4. Results

The results are organized as follows –

Part-1 - Reviewing graphical summaries / plots to get a sense of the data – including plots to visualize relationship of each of the explanatory variables with response variable.

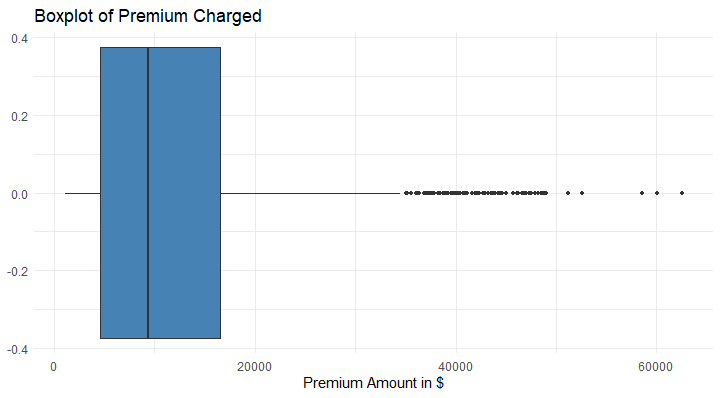
Part-2 – Fit Multiple Linear Regression(MLR) model

Part-3 – Regression Diagnostics to analyze the fit of MLR model

Part-4 – Two-sample means – analyze difference in average premium for smokers vs non-smokers

**Part-1 - Reviewing plots to get a sense of the data –**

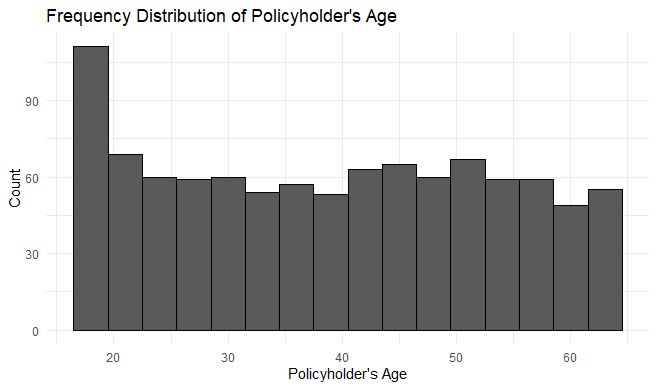
Boxplot of premium-amount –





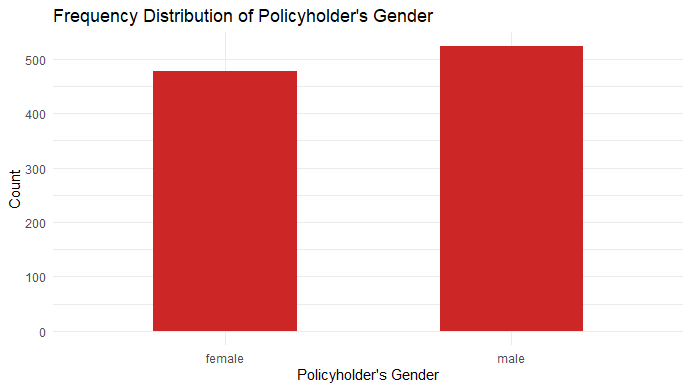
Using the IQR method, we see that there are 96 outliers for the premium-amount.

Frequency distribution of Age –

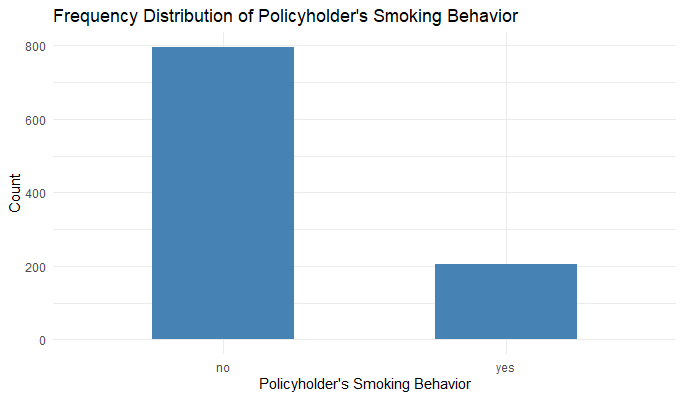


The age distribution seems to be uniform, except for a higher number of policyholders less than 20 years old.

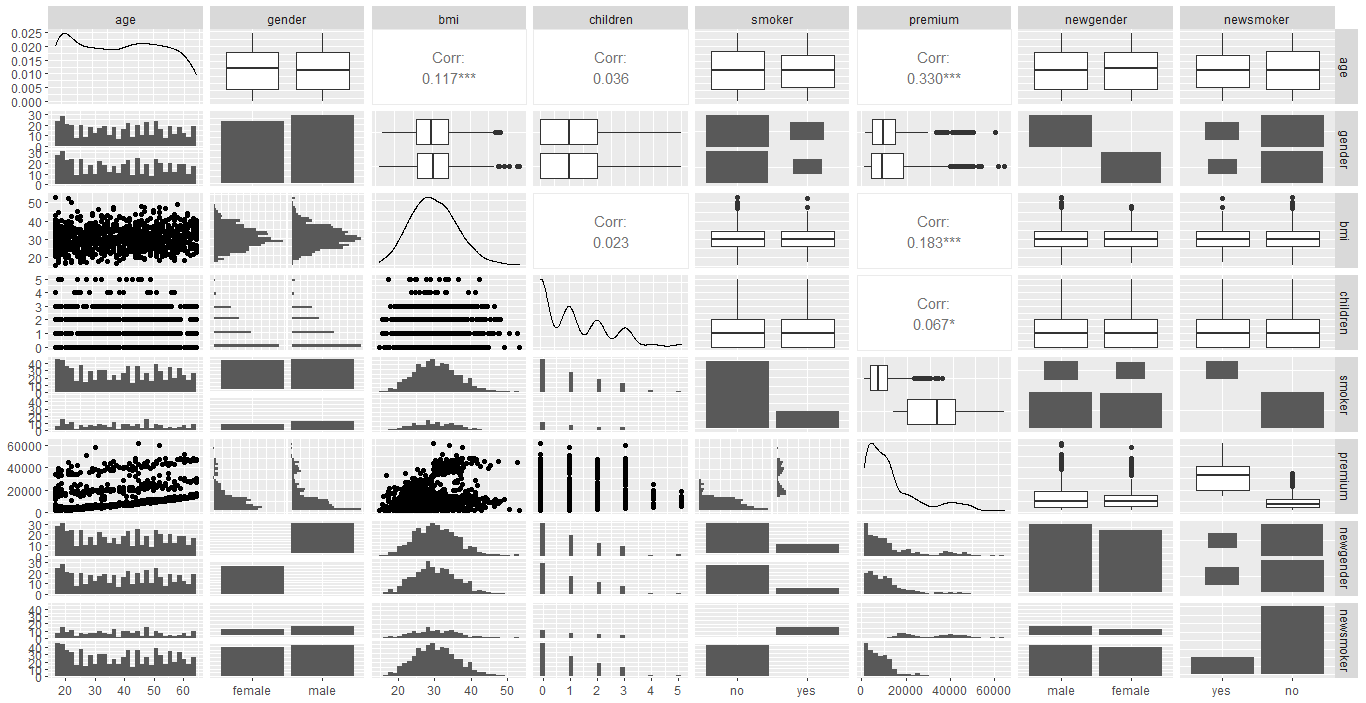
Frequency distribution of gender –



Frequency distribution of smoking behavior –



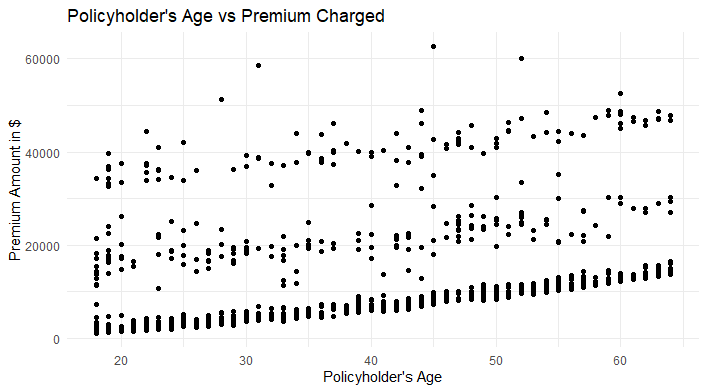
Pairwise comparison of all attributes –



Inferences - there seems to be some clustering between age & premium.

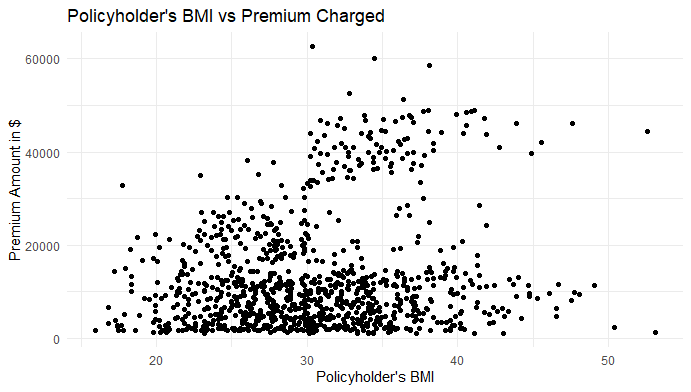
**Visualizing the relationship of each of the explanatory variables with response variable – premium-amount –**

Age vs premium –

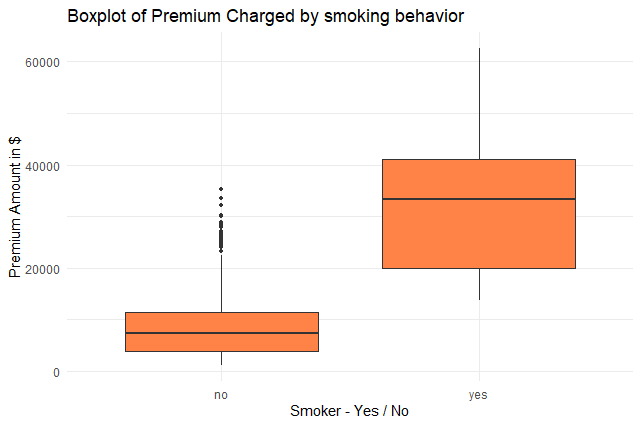


There seems to be 3 groups/clusters in premium by policyholder’s age.

BMI vs Premium -

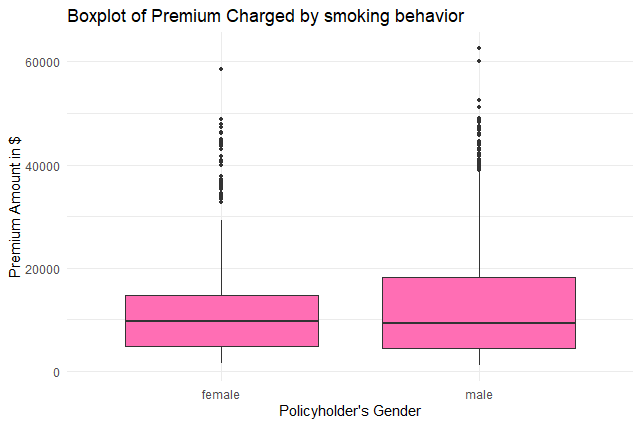


Premium by smoking behavior –



We see that smokers have a higher and a wider premium range compared non-smokers. Non-smokers have outliers that have premiums in the range of smokers.

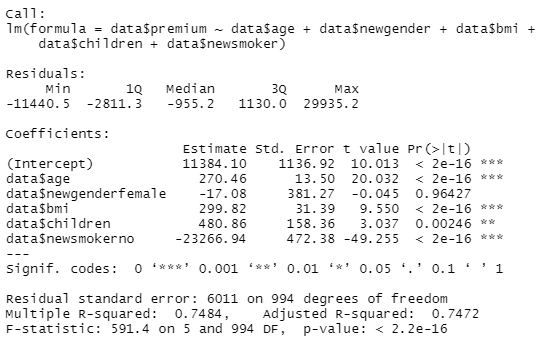
Premium by gender –



We see that men and women have about the same range of premium amounts while men have a slightly larger variance in premium amounts compared to women. Both men and women have outliers.

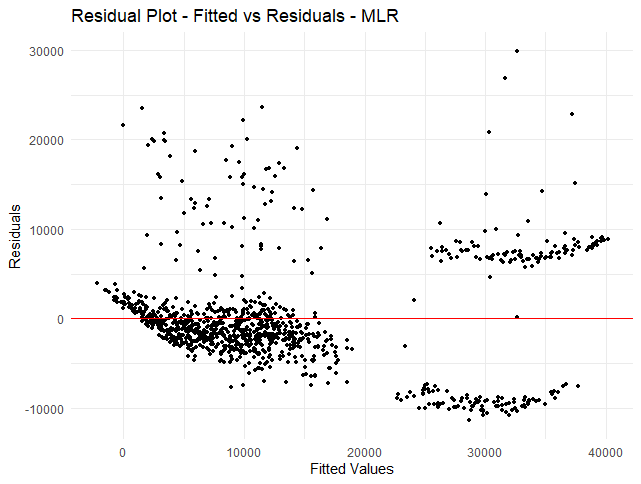
**Part-2 – Fit Multiple Linear Regression(MLR) model –**

Summary of the MLR model –



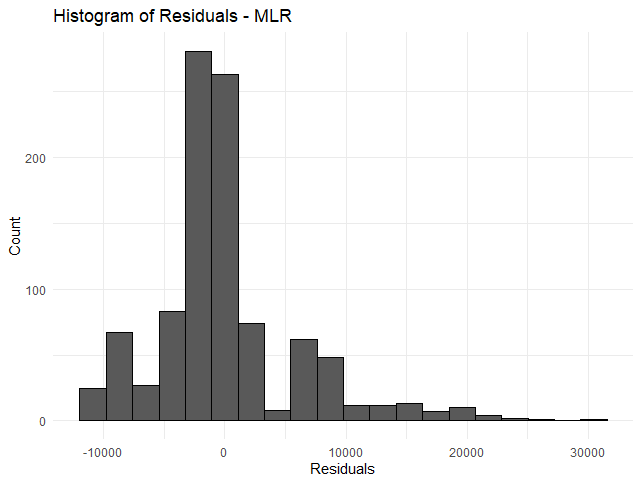
**Part-3 – Regression Diagnostics –**

Residual Plot of fitted vs residuals of the MLR model –



We see there are clusters across the residuals as well as across the fitted values.

Histogram of residuals –



Review of linear regression model assumptions –

Linearity – From the residual-plot, we see there are 3 groups/clusters in the residual-plot – this implies that the true relationship between the explanatory variables and the premium is not linear. Hence the linearity assumption is NOT met.

Independence – We assume that the insurance data collected, and the sampling process used to obtain this sample dataset has ensured that the data is independent.

Constant Variance – We see from the residual-plot that the residuals are clustered and not evenly distributed from the left to the right. This assumption is NOT met.

Normally distributed residuals – From the histogram of the residuals, we see that the distribution is right-skewed and deviates from normality. Hence this assumption is NOT met.

**I conclude that the relationship between explanatory variables(age, gender, bmi, number-of-children and smoking behavior) and premium amount is not linear.**

From the residual-plot, we see there are clusters across the residuals as well as across the fitted values. I tried removing each of the explanatory variables to understand which one of these were causing the clusters. I see that Age results in clusters in the residuals and smoking-behavior results in clusters in the fitted values.

Since the multiple linear regression model is not an appropriate fit for the data, I will not be performing the F-test or subsequent t-tests to determine the relative significance of explanatory variables in predicting response variable.

Since policyholder’s smoking behavior seems to have a high impact the premium-charged and the boxplot shows a large difference in average premium for smokers/non-smokers, I will analyze the difference in premium-amount by whether the policyholder is a smoker or not.

**Part-4 – Two-sample means – analyze difference in average premium for smokers vs non-smokers**

I have performed the two-sample test of means with the 5-step procedure to analyze difference in average premium for smokers & non-smokers -

Setup the hypothesis & select the alpha-level –

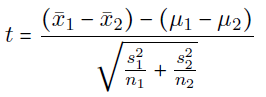
H0 / null hypothesis : mu-smokers = mu-nonsmokers i.e, average premium for smokers is same as average premium for non-smokers.

H1 / alternate hypothesis : mu-smokers != mu-nonsmokers i.e, average premium for smokers is different from average premium for non-smokers.

Alpha-level : 0.05

Select the appropriate test-statistic –

Though the sample-size is greater than 30, I will use the t-test statistic since it is considered to be more robust to non-normality.



State the decision-rule –

n1 = 204(smokers) and n2 = 796(non-smokers)

Degree of freedom df = 230.38 (from the output of t.test() function)

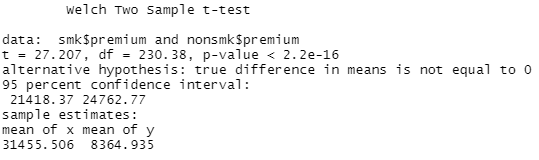
Critical value associated with right-hand tail probability alpha/2 = 0.025 is 1.970315

Therefore, we reject the null hypothesis if |t| >= 1.970315

Otherwise, we do not reject the null-hypothesis.

Compute the test statistic –

Output of t.test() function -



Conclusion –

t=27.207 > 1.970315 and p-value (2.2e-16) < alpha(0.05) – therefore, **we reject the null hypothesis and conclude that the average premium for smokers is different from the average premium for non-smokers.**

**The 95% confidence interval is (21418.37, 24762.77) i.e, we are 95% confident that, at alpha=0.05, the true difference in average premium for smokers & non-smokers is between $21418.37 & $24762.77 (p-value = 2.2e-16).**

## 5. Conclusions and limitations

Since we see the assumptions for linear regression are not met, I conclude that the relationship between explanatory variables(age, gender, bmi, number-of-children and smoking behavior) and premium amount is not linear. Therefore, any inferences drawn from this multiple linear regression model would be inaccurate and misleading. A more complex model is required to accurately model the relationship between explanatory variables & premium.

Further, we see that at alpha=0.05, we are 95% confident that the true difference in premium between smokers and non-smokers is between $21418.37 & $24762.77