Analysing of Medical charges using Multiple Linear Regression model.

Albert NDENGEYINTWALI

Kigali – March 3, 2025

Contents

Introduction

Based on HealthAffairs research

According to **(National Health Care Spending research for 2022)**, Health care spending in the US grew 4.1% which was still a faster rate of growth than the increase of 3.2% in 2021. strong Medicaid and private health insurance spending growth, including a turnaround in the net cost of insurance, was somewhat offset by continued declines in federal spending. Given this backdrop, our study aimed to investigate the factors influencing variations in medical claims or charges. To explore these variations, we applied Multiple Linear Regression (MLR) using the available data to identify key drivers.

Introduction ...

What is Multiple Linear Regression (MLR)?

- Multiple linear regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

Below is the formula of multiple linear regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

- And according to the above model we can explain this model by y is the dependent variable, each x_i is the independent variable for $i = 1, 2, \cdots, n$, each β_i is the coefficient while ϵ is the error term.

Why Multiple Linear Regression over other statistical method ?

- The reasons why we chose MLR over other methods is because of structure of the data available.
- Where we had **Multiple Independent Variables,Quantitative Data**. Which provided coefficients for each independent variable, making it easy to interpret the impact of each variable on the dependent variable.

Problem statement

- In this project, we aimed to understand how factors such as **age, BMI, sex** and **region** influenced medical insurance claims.
- To analyze this relationship, we employed a statistical technique called **Multiple Linear Regression (MLR)**, which allowed us to quantify the effect of each independent variable on claim amounts.
- The insights gained from this analysis could help insurance companies assess risk, predict claim amounts, and adjust pricing strategies accordingly.
- Our analysis would also evaluate the significance of each predictor variable to determine which factors have the most substantial impact on medical claim costs.

Methodology

The general methodology we followed based on this analysis is as follows:

- Data Preparation.
- Exploratory Data Analysis.
- Multiple Linear Regression (MLR) model Building.
- Interpretation of the model

Table showing detailed information

Table: data_mcharges_extended csv

n_k	age x_1	bmi x_2	charges y	sex X_3	regions X_4
1	19	27.9	16885	female	southwest
2	62	26.29	27809	female	east
3	48	28	23568	male	southwest
4	53	22.88	23245	female	east
5	20	22.42	14712	female	northwest
6	28	23.98	17663	male	east
⋮	⋮	⋮	⋮	⋮	⋮
118	18	21.66	14283	female	east
119	42	24.605	21259	male	east
120	29	21.85	16115	female	east
121	32	28.12	21472	male	northwest
122	30	23.655	18766	female	northwest
123	62	26.695	28101	male	east
124	61	29.07	29141	female	northwest

Explanation of variables

Dependent variable

'Charges': The medical costs incurred by the individual in $

Independent variables

1. 'Age': Age of the individual (Unit: Years).
2. 'BMI': Body Mass Index, a measure of body fat based on height and weight (Unit: kg/m^2).
3. 'Sex': The gender of the individual (categorical: Male or Female).
4. 'Region': The geographic area where the individual resides (categorical: East, Northwest, Southwest).

.

Substitution of our results in the model

The estimated Multiple Linear Regression (MLR) model is given by:

$$\text{Charges} = -7.904 + 254.341 \cdot \text{Age} + 449.777 \cdot \text{BMI} - 745.435 \cdot \text{Sex} + 25.898 \cdot \text{Region}$$

Interpretation of model

Intercept (-7.904)

- When all independent variables (Age, BMI, Sex, and Region) were zero, the predicted medical charges would be -7.904.
- Since negative medical charges were not meaningful, this suggests the intercept was just a baseline reference and should not be interpreted alone.

Age (254.341, $p < 0.001$)

- For each additional **year of age**, medical charges increase by **$254.34**, assuming all other factors remain constant.
- The p-value ($<$ **2e-16**) was very small, indicating **age was a statistically significant predictor** of medical charges.

BMI (449.777, $p < 0.001$)

- For each **unit increase in BMI**, medical charges increased by **$449.78**, keeping other variables constant.
- The p-value ($<$ **2e-16**) suggested that **BMI was a highly significant predictor** of medical charges.

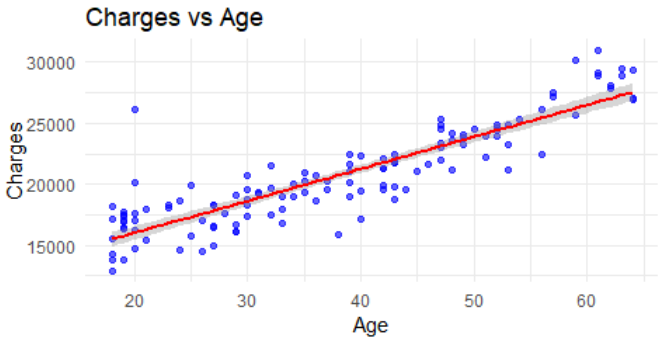
Sex (-745.435, $p = 0.003$)

- Since Sex was a categorical variable (e.g., 0 = Female, 1 = Male), the negative coefficient meant **males (coded as 1) tend to have $745.43 lower medical charges than females (coded as 0)**, holding all else constant.
- The p-value (**0.00305**) was statistically significant, meaning that **sex had a significant effect on medical charges**.

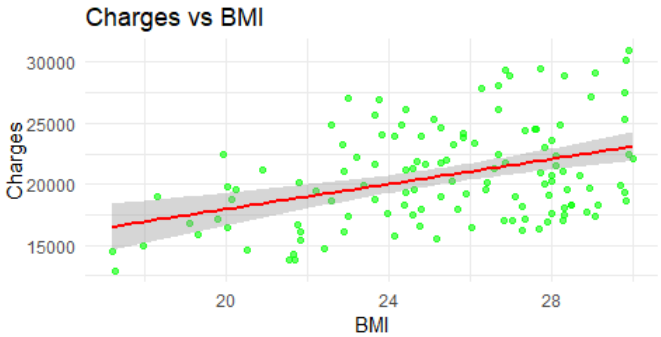
Region (25.898, $p = 0.871$)

- The coefficient was **small (25.898)** and **not statistically significant ($p = 0.871$)**.
- This meant that **region did not have a strong impact on medical charges**, and any variation in charges due to region may be due to random chance rather than a true effect.

How age or bmi affect Charges individually



- Correlation between Age and Charges (0.88): There is a strong positive correlation, meaning that as age increases, medical charges tend to increase significantly.



- Correlation between BMI and Charges (0.40): There is a moderate positive correlation, indicating that higher BMI is associated with higher medical charges, but the relationship is weaker compared to age.

Conclusion

Model Significance and Insights

- **Age and BMI were the strongest predictors** of medical charges (high coefficients and very significant p-values).
- **Sex also influences medical charges**, with females generally having higher medical costs than males.
- **Region does not significantly impact medical charges**, so it could not be a useful predictor in this model.

End

Thank You