**Raj Varshith Reddy Malgireddy**

**Understanding Factors Affecting Medical Costs**

1. **Introduction**

The rising costs of healthcare around the world are a big worry. To tackle this issue, it's important to really understand why medical expenses are going up. In a time when people are more aware of their health and healthcare systems are under a lot of pressure, it's crucial to thoroughly study the things that make medical costs go high. This project is taking a closer look at the many factors that affect how much an individual has to spend on medical care. We're especially looking at things like age, gender, body mass index (BMI), how many children someone has, whether they smoke, and differences in healthcare costs between regions.[1].

**Background and Rationale**

Healthcare costs are going up a lot, and we really need to understand why. It's not enough to just look at numbers—we need detailed insights. This is important because the high costs are affecting people's wallets, as well as the overall healthcare system and the economy. We want to figure out what's causing these high costs so that we can help doctors, insurance companies, and people who make healthcare rules to use this information and make healthcare better and cheaper.

**Project Objectives**

We're trying to figure out why medical insurance costs so much. We're looking at different things like age, gender, body weight, how many kids someone has, if they smoke, and differences between regions. By carefully studying all these factors, we want to understand how they are connected and affect the cost of healthcare. Also, we're taking a closer look at smoking because it seems to be strongly linked to high costs based on our first look at the data.

**Contextualizing the Data**

We're looking at information from 1138 people in the USA who had medical expenses in 2018. The data includes usual things like age and gender, but also other details like body weight, how many kids they have, if they smoke, and where they live. We got this data from Kaggle, and it's like our canvas to create a full picture of what factors affect how much people spend on medical care in today's world.

**Significance of Methodology**

To make sense of the data's complexity, we're using something called the Multiple Regression Model. It's a strong and easy-to-understand method for datasets with many different factors. This method helps us not only measure how different things are connected to the cost of medical insurance but also consider other things that might affect the results. This way, we can get a detailed understanding of all the factors involved.

**Research Contributions**

The results of this study are important outside of just learning things. For insurance companies, knowing what things affect medical costs helps them decide how much to charge for health insurance. Also, our look into how body weight (BMI) affects healthcare costs gives useful information for finding ways to prevent obesity in a way that saves money. [3].

This study might help find groups or people who are more likely to have big medical expenses. This information is important for making fair and effective healthcare rules. Also, it can teach patients and insurance companies about how lifestyle choices and healthcare decisions can affect money matters.

In short, this project sits at the crossroads of healthcare money matters and rules. We're thoroughly looking into the things that affect how much medical care costs. By carefully studying lots of different factors and using smart math methods, we hope to provide useful information for planning healthcare, setting insurance policies, and making rules.

1. **Literature Review**

**Healthcare Expenditure and Demographic Factors**

Lots of studies show that things like age and gender are closely connected to how much people spend on healthcare. For example, as people get older, they often have more health problems, and that means more medical costs. Also, the data includes information about gender, and research says that whether someone is a man or a woman can affect how they use healthcare and how much it costs. [5].

**Lifestyle Choices and Healthcare Costs**

Many studies focus on how the way people live affects how much they spend on healthcare. One big factor is smoking. People who smoke have a higher chance of getting sick, so they end up having to pay more for medical care than those who don't smoke. Knowing how smoking affects money matters is important for making specific plans and rules to help both individuals and the healthcare system. [6].

**Regional Disparities in Healthcare Spending**

People have been looking into why healthcare costs are different in various places. The way healthcare is set up, the rules in each area, and how rich or poor people are can make prices vary. In the data, there's a category for 'region' that lets us check out these differences. Understanding them can help us figure out how to deal with the unequal healthcare costs in different areas by making targeted plans for each region. [7].

**Methodological Landscape: Multiple Regression Model**

Using the Multiple Regression Model is in line with how researchers generally study healthcare money matters. This model helps us look at many different things at once, like age, smoking, and where someone lives, and see how they all connect to medical insurance costs. Because it looks at lots of factors together, this model helps us get a detailed understanding of what really affects how much people spend on healthcare. [8].

**Implications for Insurance Companies and Obesity Prevention Strategies**

This study can help insurance companies a lot. Figuring out what things affect medical costs is important not just for deciding how much to charge for health insurance but also for coming up with pricing plans that make sense for people's wallets and for the overall healthcare situation.

Additionally, looking into how body weight (BMI) affects healthcare costs is helpful in the ongoing fight against obesity and the related medical expenses. By figuring out how BMI is connected to overall healthcare expenses, this research brings a useful perspective to the conversation about preventing obesity. It could also provide insights for making policies that aim to lower obesity rates and the financial strain it puts on the healthcare system. [9].

**Policy Implications and Educational Aspects**

This study doesn't just help insurance companies, doctors, and policymakers right away. It also has bigger implications for making healthcare rules. Figuring out which groups or people are more likely to have big medical costs can help create plans that make sure everyone can get healthcare fairly. Plus, this research teaches us about how lifestyle choices and healthcare decisions can affect how much money we might have to spend on our health.

**Conclusion of the Literature Review**

To sum it up, the literature review lays the foundation for our current research by putting it into the bigger picture of healthcare money matters. It combines what we already know about things like age, lifestyle choices, where people live, and how we study these things. This review gets us ready to dig deep into the Multiple Regression Model, the details of the data we're using, and what we find, all of which add to the ongoing conversation about healthcare costs.

1. **Methodology**

**Hypothesis:**

The null and alternative hypotheses are usually built up for each predictor variable in a multiple regression model. The coefficients, or "parameters," of the predictors in the regression equation are connected to the hypotheses.

Two predictor variables (X1 and X2) in a multiple linear regression model may be represented by the following regression equation:

Where:

* This variable is dependent on Y.
* These are the independent variables, X1 and X2.
* The value of y-intercept is β0.
* The coefficients β1 and β2, respectively, signify the change in Y that accompanies a one-unit change in X1 and X2, respectively.
* The incorrect term is .

For every predictor, the null and alternative hypotheses are as follows:

**Regarding X1:**

The null hypothesis (H0) states that β1 = 0 (meaning that X1 does not predict Y.)

Hypothesis Alternative (Ha): β1 ≠ 0 (X1 meaningfully predicts Y)

**Regarding X2:**

β2 = 0 (X2 does not significantly predict Y) is the null hypothesis (H0).

Hypothesis Alternative (Ha): β2 ≠ 0 (X2 substantially forecasts Y)

A t-test is used to examine these assumptions. The null hypothesis is rejected, and it is determined that the predictor substantially predicts the dependent variable if the t-test p-value is less than the significance level, which is typically 0.05.

Regarding the Medical Cost dataset, the multiple regression model might like this if we considered the following factors as independents: age, sex, bmi, children, smoker, and area; charges would be the dependent variable.

A multiple regression model might be used to the Medical Cost dataset to forecast the dependent variable "charges" based on the independent variables "age," "sex," "bmi," "children," "smoker," and "region." The following would be the assumptions for each predictor in the model:

* **Regarding "age":**

The null hypothesis (H0) states that there is no connection between "age" and "charges."

An alternative hypothesis (Ha) states that "age" and "charges" are related.

* **Regarding "sex":**

The null hypothesis (H0) states that there is no connection between "sex" and "charges."

An alternative hypothesis (Ha) states that "sex" and "charges" are related.

* **Regarding "bmi":**

The null hypothesis (H0) states that there is no connection between "charges" and "bmi."

The alternative hypothesis (Ha) states that "bmi" and "charges" are related.

* **Regarding "children":**

Null Hypothesis (H0): "Children" and "charges" have no connection.

An alternative hypothesis (Ha) states that "children" and "charges" are related.

* **Regarding "smoker":**

The null hypothesis (H0) states that there is no connection between "smoker" and "charges."

Hypothesis Alternative (Ha): A correlation exists between the terms "smoker" and "charges."

* **Regarding "region":**

The null hypothesis (H0) states that there is no connection between "region" and "charges."

An alternative hypothesis (Ha) states that "region" and "charges" are related.

1. **Data used for Empirical Application**

The dataset called the Medical Cost Personal Dataset is a collection of health-related information that is commonly used in healthcare analytics and machine learning. It contains 1,339 entries. Includes important details, for understanding and predicting individual medical costs covered by health insurance. This dataset provides information like age and gender health indicators such as Body Mass Index (BMI) family related factors like the number of children or dependents and lifestyle factors such as smoking status. Additionally, it also includes information, about where the beneficiaries live in the United States.

**Features of the Dataset:**

The following characteristics are included in the dataset:

Age: The principal beneficiary's age.

Gender: The gender of an insurance contractor, either male or female.

The BMI, or body mass index, provides data on weights that are relatively high or low for a certain height. The ideal height to weight ratio of 18.5 to 24.9 is used to construct this objective body weight index (kg / m^2).

Children: Total number of dependents / number of children with health insurance coverage.

Smoker: This indicates if the recipient smokes or not.

Region: The residence region of the recipient in the United States; this may be in the northeast, southeast, southwest, or northwest.

This dataset stands out for having no missing values, which makes the process of preparing the data easier. The tilt to the right in the distribution of individual medical expenses sheds light on the financial elements of healthcare. The dataset highlights geographical differences, showing that medical costs are often greater in the Southeast than in the Southwest.

The dataset is mostly used for medical expense prediction, particularly for teaching and learning reasons. It is a useful tool on sites like Kaggle.

**4.1 Data Preparation**

A critical first step in every machine learning effort is data preparation. When getting ready to prepare the Medical Cost dataset, you may want to think about doing the following steps:

**Key Features:**

**Managing Missing Values:** Thankfully, there are no missing values in the Medical Cost dataset. If that happened, though, you would have to choose how to deal with them. Imputation, which substitutes statistical estimates such as the mean or median for missing data, and deletion, which eliminates rows or columns with missing values, are common techniques.

**Encoding Category Variables:** The dataset includes variables like "sex," "smoker," and "region." These must be transformed into a format that the machine learning system can comprehend. One popular method is one-hot encoding, which generates new binary columns for every variable category.

**Scaling Numerical Features:** Certain numerical features, such as "children," "age," and "bmi," may require scaling. The performance of many machine learning algorithms is improved by scaling numerical input variables to a standard range. This covers distance-measuring algorithms like k-nearest neighbors and algorithms that employ a weighted sum of the input like linear regression.

**Feature Engineering:** The goal of feature engineering is to develop new features that more accurately capture the underlying patterns seen in the data. One way to capture the combined influence of these factors on medical expense would be to develop an interaction term between the variables "smoker" and "age."

Table 1: Summary Statistics of the Attributes of Medical Cost

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Explanatory Variables** | **Mean/Percentage** | **Stan.Dev** | **Minimum** | **Maximum** |
| **Region**  Southeast  Southwest  Northeast  Northwest | 28%  24%  24%  24% |  |  |  |
| **Sex**  Male  Female | 51%  49% |  |  |  |
| **BMI**  < 16  16-17  17-18.5  18.5 - 25  25-30  30-35  35-40  > 40 | 0%  0%  1%  17%  29%  29%  17%  7% | 6.09 | 15.96 | 53.13 |
| **Smoker**  Yes  No | 20%  80% |  |  |  |
| **Age**  18-25  25-30  31-35  36-40  40-45  46-50  50-55  56-60  61-65 | 23%  10%  10%  9%  10%  11%  10%  9%  7% | 14.04 | 18 | 64 |
| **Children**  0  1  2  3  4  5 | 43%  24%  18%  12%  2%  1% | 1.20 | 0 | 5 |

The analysis of the dataset reveals important demographic characteristics that may influence medical charges. The distribution across regions, with Southeast being slightly more represented, suggests potential regional variations in healthcare utilization and costs. The balanced gender distribution implies that any observed differences in medical charges are less likely to be influenced by gender-specific factors. The prevalence of non-smokers among the population may contribute to a healthier overall profile, potentially impacting medical costs positively. BMI categories highlight a concentration in the 25-35 range, which could have implications for obesity-related health issues. Age diversity, with a mean age of 14.04 and a moderate standard deviation of 18, suggests a varied demographic that may experience a range of medical conditions. Family structure, as indicated by the number of children, provides insights into potential healthcare needs associated with family dynamics. Overall, the dataset's demographic composition serves as a valuable context for understanding and interpreting patterns in medical charges, offering a foundation for further analysis and exploration.

1. **Results and Discussion**

A multiple linear regression is performed to assess the impact of number of factors on the  
overall Medical Insurance Cost. The analysis of the factors influencing medical charges in the dataset reveals several key insights. First and foremost, the geographical region plays a significant role in determining medical costs. Individuals in the Southeast and Southwest regions tend to have lower medical charges compared to those in the Northwest, as indicated by the negative coefficients (-1,035.0 and -960.0). The statistical significance of these coefficients, supported by t-values and p-values, underscores the regional impact on healthcare expenses.

Table 2: Parameter Estimation Results of Multiple Linear Regression Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Explanatory Variables** | **Estimate** | **Std.Error** | **t-value** | **Pr(>|t|)** |
| **Region**  Southeast  Southwest  Northeast  Northwest | -1,035.0  -960.0  -353.0 | 478.7  477.9  476.3 | -2.162  -2.009  -0.741 | 0.030782 \*  0.044765 \*  0.458769 |
| **Sex**  Male  Female | -131.3 | 332.9 | -0.394 | 0.693348 |
| **BMI**  < 16  16-17  17-18.5  18.5 - 25  25-30  30-35  35-40  > 40 | 339.2 | 28.6 | 11.860 | < 2e-16 \*\*\* |
| **Smoker**  Yes  No | 23,848.5 | 413.1 | 57.723 | < 2e-16 \*\*\* |
| **Age**  18-25  25-30  31-35  36-40  40-45  46-50  50-55  56-60  61-65 | 23%  10%  10%  9%  10%  11%  10%  9%  7% | -10472.63  -9,839.80  -9,109.25  -8008.35  -6318.19  -4989.53  -2536.01  -1184.49  NA | 1140.47  729.03  829.02  839.14  849.12  829.98  822.61  823.61  NA | < 2e-16\*\*\*  < 2e-16 \*\*\*  < 2e-16 \*\*\*  < 2e-16 \*\*\*  5.09e-14 \*\*\*  1.71e-09 \*\*\*  0.00212 \*\*\*  0.15827 \*\*  NA |
| **Children**  0  1  2  3  4  5 | 673.63 | 144.72 | 4.655 | 3.57e-06 \*\*\* |

**\***, **\*\***, and **\*\*\*** indicate statistical significance at 10%, 5%, and 1% respectively

Note: In different categories some data are omitted because each of them serves as the reference category, and the coefficients for the reference categories are absorbed into the intercept.

Furthermore, lifestyle choices, such as smoking, emerge as a major contributor to medical charges. Smokers incur substantially higher charges, with a coefficient of 23,848.5, emphasizing the financial implications of this habit on healthcare costs. Conversely, gender does not appear to be a significant determinant, as the coefficient for male sex lacks statistical significance.

The influence of age and BMI on medical charges is evident, with both factors demonstrating statistically significant effects. Age categories exhibit varying impacts, with the 18-25 age group experiencing a significant decrease in charges compared to other groups. BMI, representing body weight, shows a positive relationship with medical costs, reinforcing the connection between health status and healthcare expenses. Additionally, the number of children is associated with higher medical charges, revealing the financial implications of family size on healthcare expenditures. This comprehensive analysis provides valuable insights for healthcare professionals and policymakers seeking to understand the multifaceted factors influencing medical costs.

1. **Conclusion and Policy Implications**

In summary, the study revealed important factors influencing medical charges. Differences in costs were observed among regions, with the Southeast and Southwest showing higher charges. Lifestyle choices, like smoking, and personal characteristics such as age, BMI, and having children also played a role in determining healthcare expenses. Policymakers should consider these factors when creating strategies to make healthcare more affordable and accessible.

Based on the findings, it's recommended to focus on reducing regional disparities in healthcare costs. Targeted campaigns to promote healthier lifestyles, especially in regions with higher charges, could be beneficial. Additionally, policies supporting preventive care tailored to different age groups and BMI ranges might help in managing healthcare costs more effectively.

For future research, exploring specific health conditions and their impact on charges, considering how various factors interact, and including socioeconomic aspects in the analysis could provide more comprehensive insights. The study forms a foundation for further research aimed at creating evidence-based policies to ensure fair and affordable healthcare for everyone.

1. **References**

[1] NCBI (2018). High-quality health systems in the Sustainable Development Goals era: time for a revolution, [Lancet Glob Health.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7734391/)

[2] Kaggle. (2017). Medical Cost Personal Datasets [[https://www.kaggle.com/datasets/](https://www.kaggle.com/datasets/mirichoi0218/insurance/data)

[mirichoi0218/insurance/data](https://www.kaggle.com/datasets/mirichoi0218/insurance/data)]

[3] [Hadi Kharrazi](https://pubmed.ncbi.nlm.nih.gov/?term=Kharrazi%20H%5BAuthor%5D), MD PhD, [Hsien-Yen Chang](https://pubmed.ncbi.nlm.nih.gov/?term=Chang%20HY%5BAuthor%5D), PhD. Assessing the Impact of Body Mass Index Information on the Performance of Risk Adjustment Models in Predicting Healthcare Costs and Utilization. NCBI (2019)

[4] [Jie Chen](https://pubmed.ncbi.nlm.nih.gov/?term=Chen%20J%5BAuthor%5D), 1 [Meizhen Zhao](https://pubmed.ncbi.nlm.nih.gov/?term=Zhao%20M%5BAuthor%5D), 2 [Renyi Zhou](https://pubmed.ncbi.nlm.nih.gov/?term=Zhou%20R%5BAuthor%5D), (2023). How heavy is the medical expense burden among the older adults and what are the contributing factors? A literature review and problem-based analysis

[5] [Agostina J. Larrazabal](https://www.pnas.org/doi/10.1073/pnas.1919012117#con1), [Nicolás Nieto](https://www.pnas.org/doi/10.1073/pnas.1919012117#con2), [Victoria Peterson](https://www.pnas.org/doi/10.1073/pnas.1919012117#con3) . (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. PNAS

[6] HHS Press Office (2023). First-of-its-Kind Study Examines Disparities in Healthcare Costs and Outcomes of Cigarette Smoking in the United States

. HHS.gov

[7] [Tom Rosenthal](https://www.researchgate.net/profile/Tom-Rosenthal?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19), [University of California, Los Angeles](https://www.researchgate.net/institution/University_of_California_Los_Angeles?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19) (2011). Geographic Variation in Health Care. Researchgate.

[8] Jason Abaluck Mauricio M. Caceres Bravo (2020). Mortality effects and choice across private health insurance plans. National Bureau of Economic Research.

[9] Zachary J. Ward ,Sara N. Bleich, (2021). Association of body mass index with health care expenditures in the United States by age and sex. Plos One.