Research On the Factors That Affect the Pricing of Healthcare Insurance

Zewei Tang and Xiaoxuan Wang *

Beijing International Bilingual Academy, Tianjin campus, Tianjin, 101300, China

* Corresponding Author Email: 1712020201@stu.hrbust.edu.com

Abstract. This essay delves into the intricate relationship between various factors and healthcare insurance charges, employing two sophisticated mathematical-statistical models: machine learning such as linear regression and skewness and kurtosis. The dataset utilized in this analysis is sourced from the reputable data science platform, Kaggle. The findings of this study indicate that several key factors play a crucial role in determining health insurance pricing. Specifically, age and smoking status are identified as significant influencers. However, the BMI (Body Mass Index) cannot affect the charges. As age increases, so do healthcare charges, reflecting the higher likelihood of health issues in older individuals. Similarly, smokers are charged more due to the increased health risks associated with smoking. Additionally, higher BMI values are not linked to higher insurance charges. Moreover, the study highlights a direct correlation between these factors and healthcare insurance pricing. As age and smoking status increase, there is a corresponding increase in insurance charges. This underscores the importance of these factors in determining the cost of healthcare insurance.

Keywords: Insurance; influencing factors; machine learning.

1. Introduction

Healthcare insurance is classified as a good that possesses the characteristic of reimbursement, which means the residents can obtain the support of the fund if individuals encounter a range of accidents and have to spend the expenditure on treatment. Nevertheless, despite the law about healthcare insurance has been modified and enhanced in recent years, the operation of the system about healthcare still has some troubles. On a hand, the resources of healthcare cannot be consumed efficiently and effectively, which leads to the economic burdens of the people, who accept the curing, aggravate significantly. On the other hand, according to the reference essay, if the individuals who ought to provide the service of the treatment and the medicines forfeit control, the existence of the healthcare insurance may lead to the price of the medical curing soars and the cost of the treatment rocket remarkably with the large frequency [1]. In addition, according to the WHO annual report in 2021, Global spending on health more than doubled in real terms over the past two decades, reaching US$ 8.5 trillion in 2019, or 9.8% of global GDP [2]. Because of the rising expense of quality healthcare, increased life expectancy, and the epidemiological shift toward non-communicable diseases, health insurance is becoming an essential commodity for everyone. Insurance data has increased dramatically in the last decade, and carriers now have access to it [3].

At a time when people are inextricably linked to insurance, it is essential to the factors to search the influences of healthcare insurance and the pricing of health insurance on that basis, as this can affect the daily consumption of the population. Therefore, predicting the reasonable price of the insurance which can relieve the economic burden of the people is not only the top priority of the approaches that enhance the welfare of the treatment but also ensure the insurance companies obtain positive earnings. The factors that influence healthcare insurance pricing are also essential projects which need to be researched.

There are plenty of factors that have influenced healthcare insurance, as it is a complex system. The foreign researcher, namely Wright, has found that healthcare insurance pricing is related to the structures of the healthcare markets [4]. Moreover, Bryce illustrated that health insurance coverage, the availability of pharmaceuticals, and hospital billing practices are factors that affect the level of access to care [5]. However, the healthcare sector produces a very large amount of data related to

patients, diseases, and diagnosis, but since it has not been analyzed properly, it does not provide the significance that it holds along with the patient healthcare costs [6]. Therefore, this essay primarily selects 3 factors (age, whether the insurers are smokers, and the charges that the insurance company spends on the insurers) and research whether the healthcare insurance pricing is affected by them. And exploring their relevance to insurance pricing through a range of appropriate mathematical models.

Regarding to the method of the research, Duan introduced multiple regression models into health insurance premium measurement and proposed a four-part model to predict outpatient and inpatient utilization and costs [7]. In addition, Tang used the kernel density estimation method in a non-parametric model to estimate the density of the distribution of the logarithm of the cost per hospitalization and used it to further estimate the mean and variance of the cost per hospitalization [8]. Moreover, some other foreign scholars use the Pareto distribution model to classify the pricing of large health insurance policies [9]. And they also claimed due to incapacity are also analyzed using a multilevel model Combined with the concept of transfer intensity, the distributions of probability durations and initial claim rates are analyzed to provide a basis for pricing incapacity insurance [10].

In conclusion, after consideration and optimization, this essay will determine whether to use the model of linear regression as well as the skewness and kurtosis to delve into the relationship between these factors and the pricing of healthcare insurance.

2. Methods

2.1. Data Source

The dataset that exists in this essay contains 1338 rows of insured data, where the Insurance charges are given against the following attributes of the insured: Age, Sex, BMI, Number of Children, Smoker, and Region. There are no missing or undefined values in the dataset. And this dataset is acquired from the KAGGLE.

2.2. Variable Introduction

This essay will select 6 indicators including age, sex, BMI, children, smoking, and region, and analyze the relationship of these six indicators and the charge. It is assumed that the age correlates with the charge. When the age increases, the charge of the healthcare insurance will also increase. That is because Increasing age will lead to the body's functions decreasing and create a variety of diseases and so that increases the charge of the healthcare insurance. And the BMI will also correlate to the charge because when the BMI rises, it represents an increase in the body's obesity level and thus an increase in the chances of getting sick. Besides it is also assumed that whether the insurers are smokers will also correlate to the charge, since smoking causes harmful substances to enter the body which increases the risk of disease. Therefore, this essay will choose these indicators to delve into the factors that have the influenced the pricing of healthcare insurance.

Table 1. Part of dataset of healthcare insurance

age	sex	BMI	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855

In the part of the dataset in table 1, the age means the age of the primary beneficiary, the sex refers to the insurance contractor’s gender, and the BMI means the body mass index, which provides the understanding that the body and weight are relatively high or low to the height. The children refer to the number of children covered by the healthcare insurance; the smokers refer to whether the person

covered by an insurance policy smokes the cigarette. The region means the beneficiary's residential area in the US. The charges are the meaning of Individual medical costs billed by health insurance. The last one "charge" refers to the cost imposed on the companies that engage in healthcare insurance, and it also means the charge that the insurers obtain from the firms of the healthcare insurance.

2.3. Method Introduction

This essay will use linear regression as well as skewness and Kurtosis to explore the factors that affect the pricing of healthcare insurance. Linear regression is one of the most basic and commonly used algorithms in machine learning, and it is the process of finding a straight line that makes this line as close as possible to all data points. This straight line can be used to predict the value of a dependent variable (target variable) based on one or more independent variables (features). The core idea of linear regression is to find a line of best fit that minimizes the total error between the predicted and actual values. Besides, the skewness essentially is used to calculate in descriptive statistics that characterizes the asymmetry of a data distribution, while the Kurtosis is the statistical measure that quantifies the shape of a probability distribution. The Kurtosis provides information about the tails and the peak of the distribution compared to a normal distribution.

3. Resulst and Discussion

3.1. Descritive Analysis

The result of the machine learning model is discussed in this section, this essay will analyze the feature vs charge for data visualization.

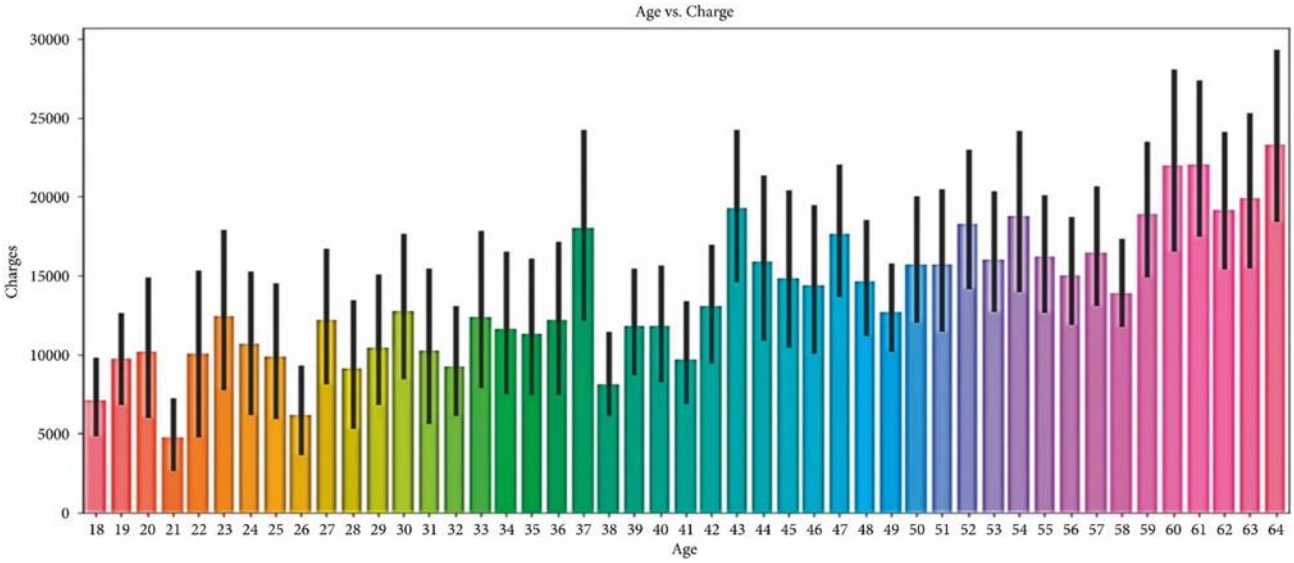

Fig. 1 Bar chart illustrating numbers of charges for different ages

Figure 1 illustrates the insurers gaining how much charge at different ages. In this diagram, the x-axis represents the age, and the y-axis represents the charge. And it can be observed that as the age increases, the charge will also rise gradually. For instance, when individuals are at the age of 18, they merely obtain a charge of approximately 7500. However, the insurers will gain the charge of nearly 22500, which means they get the highest charge of healthcare insurance when they are 64.

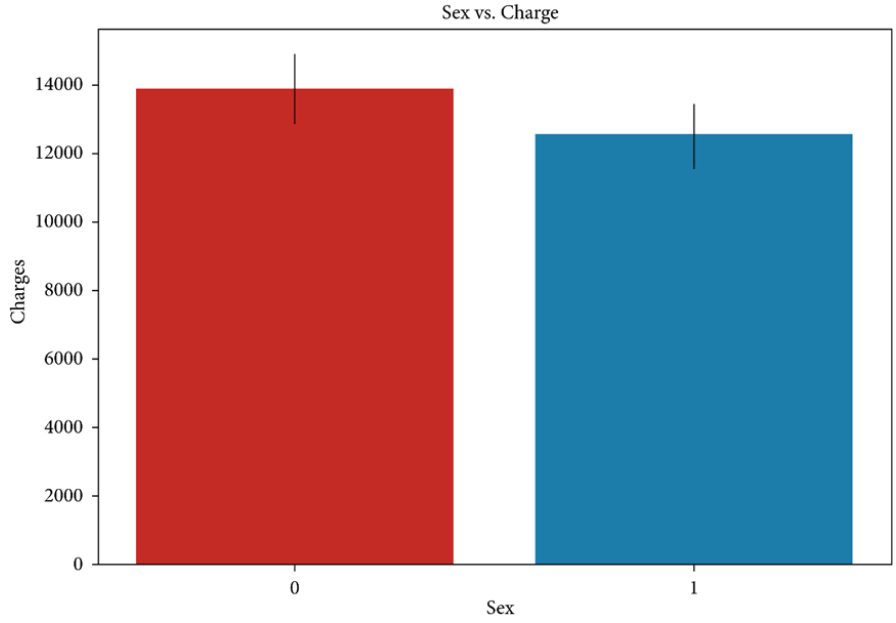

Fig. 2 Box plot illustrating numbers of charges for males and females

The above Figure 2 demonstrates the relationship between the sex and the charge. In this diagram, the x-axis means the sex, and 0 refers to the female, as well as 1 means the male. Besides, the y-axis means the charge. Thus, it can be observed that the healthcare insurance for women is relatively greater than the men, which accounts for 14000 and 13000 relatively. However, it is not a primary factor that influences the pricing.

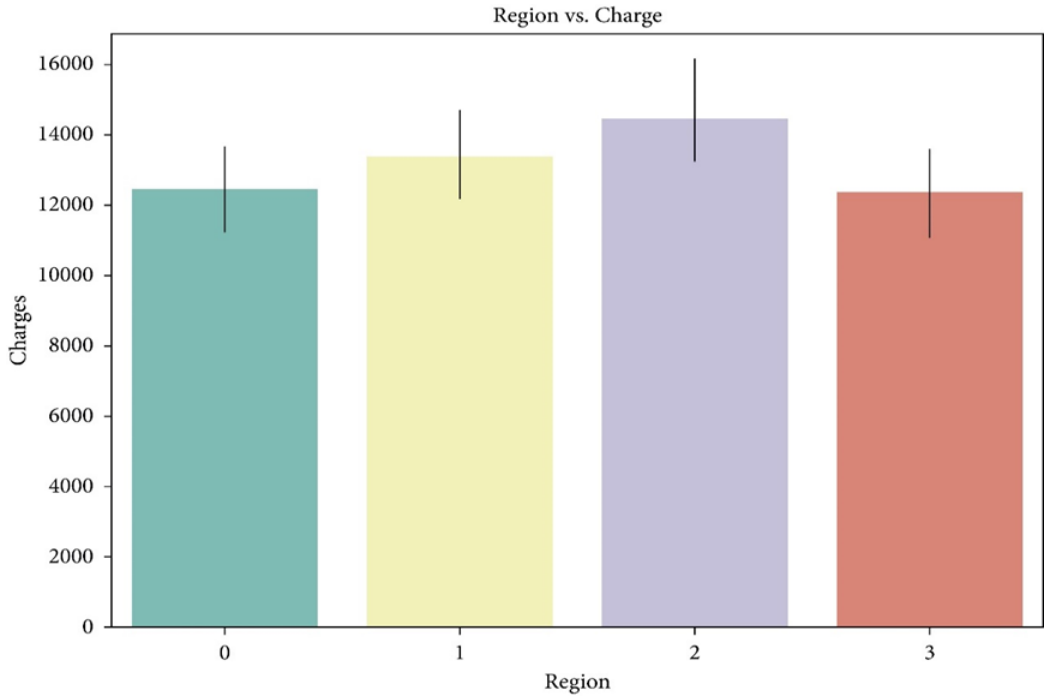

Fig. 3 Bar chart illustrating numbers of charges in different regions

Figure 3 above illustrates the relationship between the region in which the insurers live and the charge. In this case, the x-axis represents the various regions, 0 refers to the northeast, 1 means northwest, 2 means the southeast, and 3 refers to the southwest. In this figure, it can be shown that the charge on the people who live in the southeast is a little more enormous than other regions, which makes up approximately 14500. And the charge for other regions is almost constant. Thus, the regions are not the top priority of the factors.

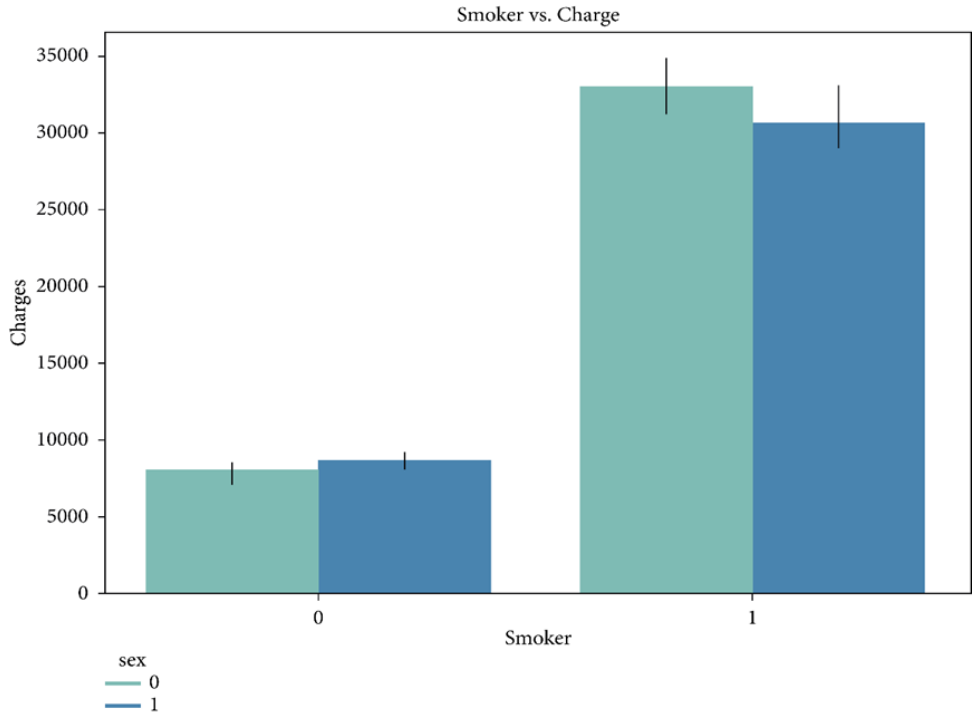

Fig. 4 Bar chart illustrating numbers of charges

The relationship between the smoker and the charge of the healthcare insurance which the insurer ought to spend is demonstrated in Figure 4. In this case, the x-axis displays the gender of the smoker, while the y-axis shows the charge of the insurance. It can be seen that the charge for the male is much more massive than the amount of the charge for the female, which accounts for 34000 and 8000 respectively at the beginning. That may be because compared with the less attraction of cigarettes for women, the cigarette hypnotizes the addicted men who have fancy for smoking every day. This picture also shows that as the habit of smoking cigarettes soars, the charge of healthcare insurance for males will plummet slightly by approximately 3000. In contrast, the amount of the charge for females increases minimally from 8000 to 8500.

3.2. Variable Distribution

The result of skewness and the Kurtosis is discussed in the below section, this essay will analyze the feature vs charge for data visualization.

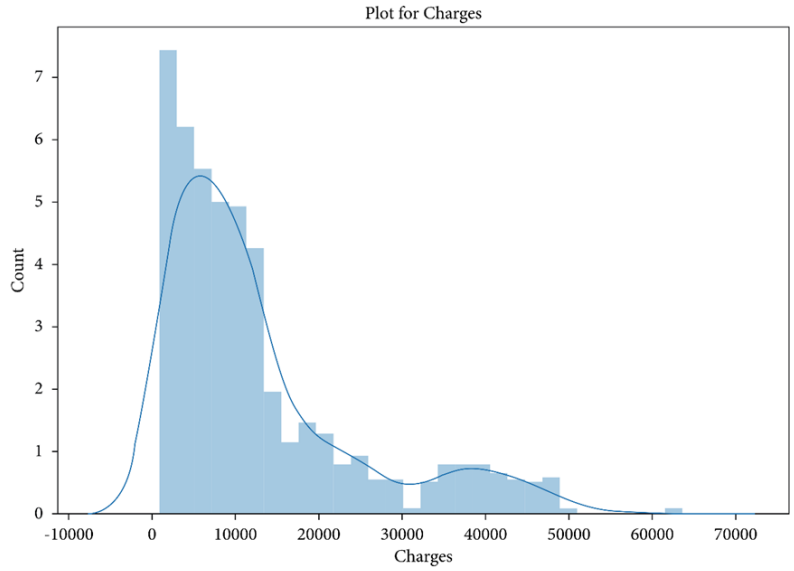

Fig. 5 Plot skew and kurtosis value for charges

The Figure 5 cannot display some outliers. In addition, it is positively skewed or a right-skewed distribution, because it has a long right tail. Besides, this diagram makes the Positively Skewed Distribution, a category of distribution, have the positive mean, median, and mode of the distribution instead of negative or zero once. In this case, the mean is greater than the median, as well as the median is greater than the mode. Besides, the skew value of the charges of the healthcare insurance can be determined by the equation $\frac{\text{mean}-\text{mode}}{\text{standard deviation}}$.

The value of the charge is 1.516. in addition, the diagram 3.6 demonstrates the platykurtic of the charge, a type of the Kurtosis, and the value of the Kurtosis is 1.606. That means the platykurtic distribution is flatter which means less peaked when compared with the normal distribution.

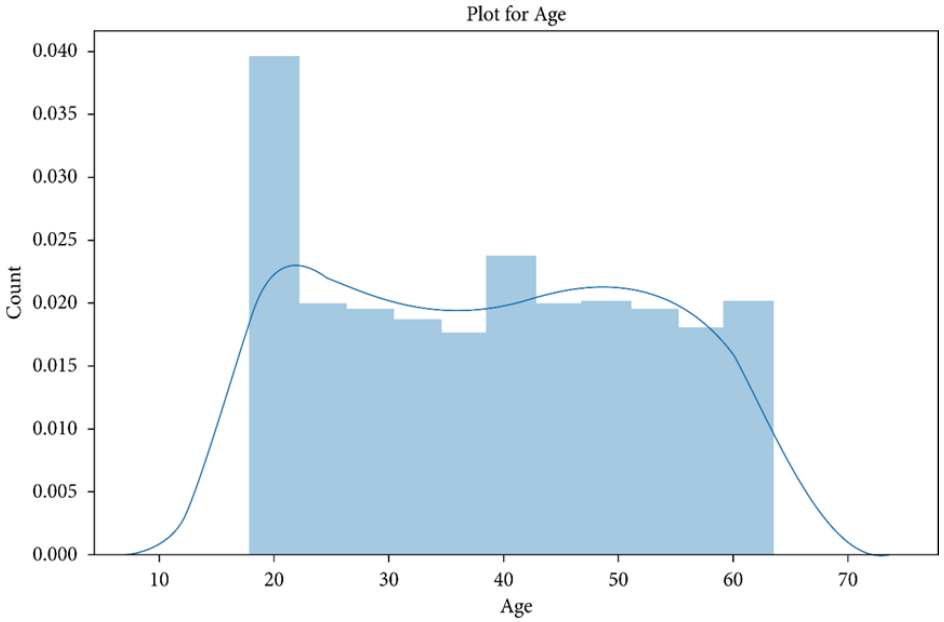

Fig. 6 Plot skew and kurtosis value for ages

The Figure 6 illustrates the values of the skewness of the age is 0.056, and it accurately shows the range of the correlation values. Besides, the kurtosis value of the age is -1.245.

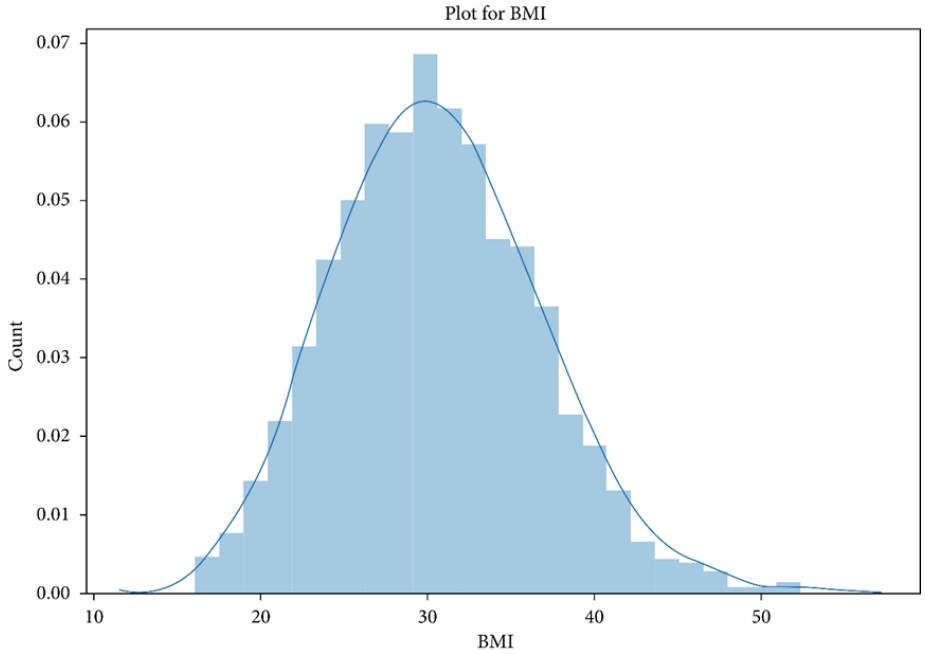

Fig. 7 Plot skew and kurtosis value for BMI

The Figure 7 illustrates the skewness and the Kurtosis of the BMI values. The diagram is similar to the normal distribution because both the left-hand side and the right-hand side have nearly equal

numbers of observations. Besides, the skewness value of the BMI is 0.284, which means the mean in this case is a little greater than the mode. In addition, the kurtosis value of the BMI value is -0.051.

3.3. Model Results

Figure 8 illustrates the connection between ages and the charges of healthcare insurance. It is observed that when the age increases, the charge will also increase. They demonstrate a linear relationship. In addition, the Adjusted R-squared which can be calculated by the equation below is 0.6991. therefore, it significantly conforms to the linear regression.

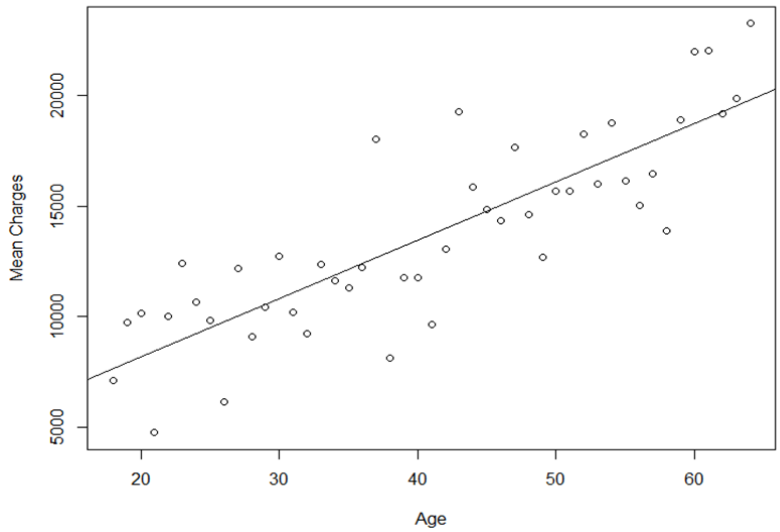

Fig. 8 Linear regression of age to charges

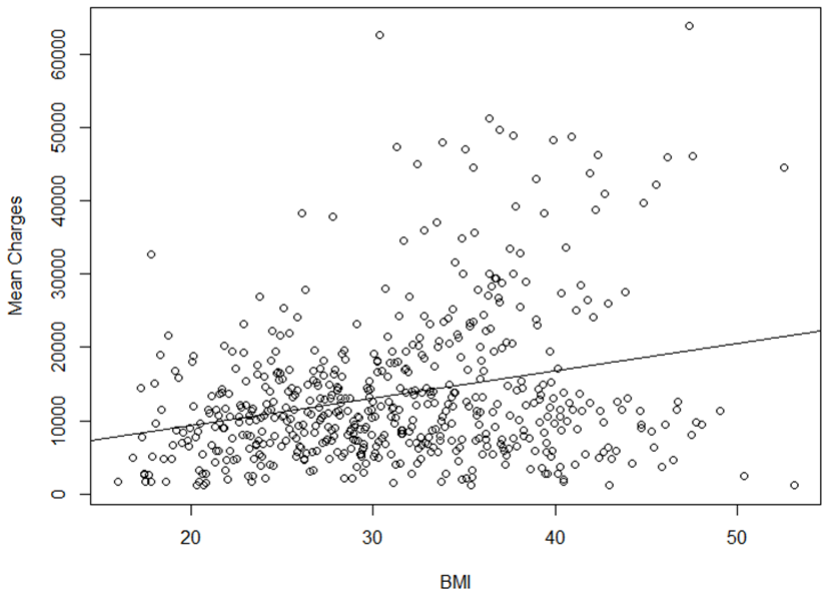

Fig. 9 Linear regression of BMI to charges

Figure 9 illustrates the relationship between the BMI values and the charges of healthcare insurance. It is seen clearly that the BMI values and the charges do not conform linear relationship, which means the increase in the BMI value is not able to lead to the growth trend of the charge. That may be because the BMI values cannot reflect the extent of the health of the body in individuals.

Besides, the Adjusted R-squared which is used to represent the relationship between the BMI and the charge is 0.06953. Thus, BMI values are not a primary factor in the charges.

By analyzing the methods of the machine learning model including linear regression and the skewness and the Kurtosis, it can be seen that the linear regression presents the results more clearly and is more accurate compared to skewness and the Kurtosis. Besides, the skewness and the Kurtosis are not able to eliminate the outliers, and the outliers are contained in the result of the skewness and Kurtosis. In addition, the linear regression can demonstrate clearly and precisely the relationship between the features that this essay determines to focus on and the charges of healthcare insurance as well. Thus, the approach of the machine learning model is more excellent than skewness and Kurtosis.

4. Conclusion

By using a machine learning model involving linear regression as well as the skewness and the Kurtosis, the analysis concludes that age, and whether or not one is a smoker has a tighter effect on health insurance pricing. Nevertheless, the BMI values that the individuals possess unnecessarily lead to health-threatening obesity which will cause the diseases with massive frequency. Thus, it cannot influence the charges in strict linearity. When the insurers are smokers, the healthcare insurance pricing will be higher. And when the age of the insurer is higher, the healthcare insurance will also be higher. Besides, as the BMI values that the people gain increase, the charges of the healthcare insurance may be unchanged. In addition, the current study also has some drawbacks, such as the small sample size and the fact that only two mathematical models were used to explore the factors influencing health insurance pricing. If there had been more time, more mathematical models would have been used to examine the influencing factors.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Zhang Ping, Xu Bing, Gan Xiaoqing. Research on the relationship between market structure, medical insurance, and medical expenses. Journal of Management Engineering, 2018, 2: 53-58.

[2] Wright D. Insurance and monopoly power in a mixed private/public hospital system. Economics Record, 2006, 82: 460-468.

[3] Sommers D. Health insurance coverage: what comes after the ACA. Health Affairs, 2020, 502-508.

[4] Duan N, et al. A comparison of alternative models for the demand for medical care. Journal of business & economic statistics, 1983, 1(2): 115- 126.

[5] Tang Guoquan, Li Rongmin, Zhang Qungui, Chen Weihua. Application of kernel estimation in the study of medical cost distribution. Journal of Applied Mathematics and Computational Mathematics, 2005.

[6] Cebrián A C, Denuit M, Lambert P. Generalized Pareto fit to the Society of Actuaries' large claims database. North American Actuarial Journal, 2003, 7(3): 18-36.

[7] Renshaw A E, Haberman S. On the graduations associated with a multiple state model for permanent health insurance. Insurance: Mathematics and Economics, 2017, 17(1): 1-17.

[8] Lv Xiaoning. Deposit Insurance Pricing Method Considering Bank Asset Liability Structure. Operations Research and Management, 2023, 10: 198-204.

[9] Wang Xiufeng, Han Hao, Liang Longyue. A Study on the Pricing of Climate Catastrophe Risk Bonds: A Case Study of Floods in Guizhou Province. Friends of Accounting, 2022, 8.

[10] Shen Jianing, Wang Fang, Huang Yingjun, et al. Pricing of Natural Rubber County Income Insurance in Hainan Province under Comprehensive Risk Zoning. Forestry Economics, 2023, 5-29.