Transactions on Economics, Business and Management Research
ISSN: 2960-1789, eISSN: 2960-2254 | Volume 10
MEEA 2024

Research on Factors Influencing Medical Insurance Cost in the US

Quanjing He *

School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, 310000, China

* Corresponding Author Email: 21064011@suibe.edu.cn

Abstract. After the COVID-19 epidemic, the growing concern over medical insurance cost and the factors that influence it are increasingly coming to the forefront. These factors include age, sex, body mass index, number of children, smoker or non-smoker and region. Therefore, this paper analyzes the impact of six factors on medical insurance costs by utilizing data on medical insurance expenses of 2,772 individuals in the US. The analysis employs a multiple linear regression model and the stepwise regression method. It is concluded that age, body mass index, number of children, smoker or non-smoker and region have a significant impact on medical insurance cost with the largest regression coefficients for age, body mass index and smoker or non-smoker. This shows that it is physiological factors that have the greatest impact and therefore it is recommended that people should better maintain good health in order to reduce medical insurance cost and save money.

Keywords: Medical insurance cost; multiple linear regression; stepwise regression.

1. Introduction

Medical insurance has always been a topic of great interest. Over the course of American history, the health care system has undergone significant transformation, becoming a highly complex sector that affects almost every element of society, which has resulted in both advancements and significant obstacles in public health [1]. Medical insurance is an important part of the US health care system. That's why it's important to study it. Moreover, Emon et al. concluded that since the start of the COVID-19 epidemic, there has been a noticeable increase in healthcare costs, so medical insurance has become a prominent area of research [2]. Medical insurance cost forecasting is a difficult endeavor because of the ever-changing nature of healthcare, the complexities and sometimes chaotic nature of data, unpredictable human behavior, fluctuations in the economy, unforeseen events, and widespread concerns about privacy [2]. Therefore, understanding the factors that affect medical insurance cost can help predict that cost. This paper aims to identify the factors affecting medical insurance cost in the US, so as to help patients have more informed choices in the process of medical treatment, to allow medical institutions and insurance companies to better manage and utilize medical insurance funds, and to allow policy makers to more accurately grasp the operation of the medical insurance system.

The factors affecting medical insurance cost are varied. Li applied the Dagum Gini coefficient method to a sample of 2010-2019 data from 31 regions in China and found that there were regional differences in medical insurance fund expenditures [3]. Mathur et al. conducted an online questionnaire survey of residents in the Lucknow region of India, and concluded that age, dependent family members, medical expenditure, health status, individual's product perception and medical insurance subscription were significantly correlated [4]. Selamat et al. searched for articles published on six major search engines from 2013-2018 on factors influencing Malaysians' demand for or willingness to pay for medical insurance, and found that the higher the level of education, the younger the age, and the more knowledgeable the person is, the higher the demand for medical insurance [5]. Sanjaya and Zen considered that decisions to purchase medical insurance are strongly influenced by demographic criteria such as income level, assets, educational background, financial dependency, and health status [6]. Therefore, this paper focuses on 6 factors (age, sex, body mass index, number of children, smoker or non-smoker, region) to study whether they have any effect on medical insurance

cost, and choose a suitable model to investigate the extent of association between these factors and the cost.

In terms of methods, Zheng empirically analyzed the factors that significantly affect the level of medical insurance fund expenditures by collecting dynamic panel data from 31 regions in China over the period 2010-2019, using dynamic panel model [7]. The model is robust and valid, but the demographic and socio-economic development level factors considered in this study are too ambitious for individuals to make decisions. Zhang, on the other hand, used a mediated effects model to study the impact of environmental pollution on health insurance expenditures and its mechanism of action [8]. The study found that environmental pollution significantly increased medical insurance expenditures, and that this effect existed in both eastern and central-western China with significant regional heterogeneity, and that the effect of environmental pollution on medical insurance expenditures was largely greater in the eastern region than in the central-western region [8]. The regional heterogeneity also provides ideas for the region variable in this paper. Wang et al. used multinomial logistic regression model to examine the effects of economic and environmental factors on medical insurance purchase decision priorities using individual-level insurance purchase microdata [9]. The scope considered in this model is more comprehensive, taking into account both objective factors in society and individual differences. Zhou and Zhao also used logistic regression model, concluding that consumers who use the WeChat Health pedometer feature longer have lower medical insurance rates [10].

In summary, after consideration and optimization, this paper will use six factors and employ multiple linear regression to examine their effects on medical insurance cost.

2. Methods

2.1. Data Source

The data for this paper is collected from the Kaggle website (Medical Insurance Cost Prediction), which was collected by M RAHUL VYAS and published in March 2024 for 2772 individuals. The original dataset remained in csv format.

2.2. Variable Selection

The 2772 people in the data used in this paper all had some level of medical insurance spending. They are between the ages of 18 and 64. The data consists of six independent variables and one dependent variable. The independent variables are Age, Sex, BMI, Children, Smoker, and Region, and the dependent variable is Charges. The names and meanings of the variables are listed in Table 1.

Table 1. List of Variables.

Variable	Logogram	Meaning
Age	x_1	Age of the person
Sex	x_2	Gender of the person Male (1), Female (0)
BMI	x_3	Body Mass Index
Children	x_4	Number of children
Smoker	x_5	Smoker or Non-smoker Yes (1), No (0)
Region	x_6	Northwest (1), Northeast (2), Southeast (3), Southwest (4)
Charges	Y	Medical Insurance Cost

To standardize the units of these independent variables, the paper performs Z-Score normalization on each independent variable to obtain new variables, z_1, z_2, z_3, z_4, z_5 and z_6.

2.3. Method Introduction

This paper uses the multiple linear regression model, which is a statistical technique used to model the relationship between a single dependent variable and multiple independent variables.

The first step is to implement a descriptive statistical analysis. Descriptive statistical analysis refers to the process of summarizing and describing the features of a dataset, such as central tendency, dispersion, and shape, through numerical or graphical methods.

The second step is to conduct a correlation analysis, which involves examining the relationships between the independent variables and between each independent variable and the dependent variable. It helps to assess the strength of these relationships, which informs variable selection and the understanding of potential multicollinearity issues.

The third step is to perform parameter estimation, which is the process of determining the values of unknown parameters in the multiple linear regression model. In this study, Ordinary Least Squares (OLS) is employed for parameter estimation, achieved by minimizing the sum of the squares of the differences between the observed and predicted values. Stepwise regression screening of variables is also used in this process, which is a method of fitting regression models in which the choice of predictive variables is carried out automatically. It involves iteratively adding or removing variables based on certain criteria, such as p-values, until no further improvements can be made according to the chosen stopping rule.

The last step is to evaluate the model. The measurement of the model fit., the significance of the regression relationship, multicollinearity testing, and residual analysis are essential tasks in this step.

3. Results and Discussion

3.1. Descriptive Analysis

Histograms are plotted for numeric variables as follows:

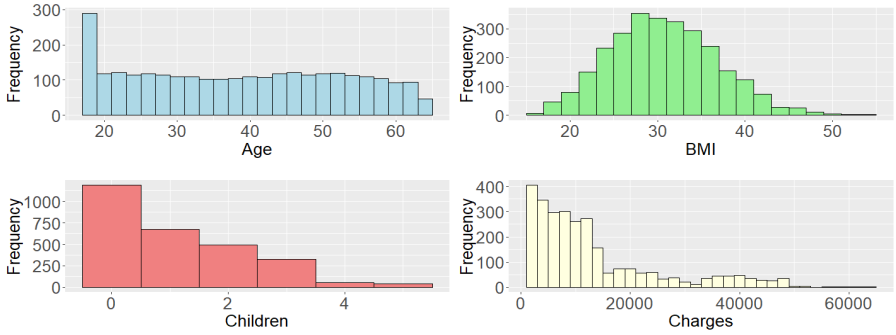

Figure 1. Histograms of Age, BMI, Children and Charges.

From Figure 1, it can be seen that when Age is greater than 20, the distribution is approximately uniform, and BMI has a two-tailed character. Both Children and Charges decrease in frequency with increasing value.

Pie charts are plotted for categorical variables, as shown below:

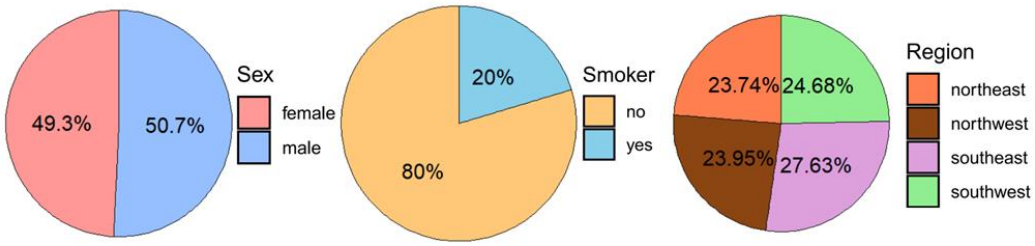

Figure 2. Pie Charts of Smoker, Sex and Region.

From Figure 2, It is easy to find far more smokers than non-smokers. The ratio of men to women is close to 1 and about the same number of people came from all four directions of the United States.

3.2. Correlation Analysis

Correlation analysis in multiple linear regression identifies relationships between variables, helping to detect multicollinearity and guide model refinement. The Pearson correlation coefficients between the variables are shown in the table below:

Table 2. Pearson Correlation Coefficients.

	Age	Sex	BMI	Children	Smoker	Region	Charges
Age	1						
Sex	-0.03	1					
BMI	0.11	0.04	1				
Children	0.04	0.02	-0.00	1			
Smoker	-0.02	0.08	0.01	0.01	1		
Region	0.00	0.01	-0.00	-0.00	0.01	1	
Charges	0.30	0.06	0.20	0.07	0.79	0.00	1

From Table 2, the correlation coefficients between the independent variables are close to 0, indicating that these independent variables are almost independent and are unlikely to have multicollinearity problems. As for the correlation coefficients between the independent and dependent variables, most notably, the correlation coefficient between Smoker and Charges is large, indicating a strong positive correlation. However, the correlation coefficients between the other independent variables and the dependent variable are relatively small.

3.3. Parameter Estimation and Model Evaluation

After analyzing the Pearson correlation coefficients of various factors, multiple regression analysis was conducted. The general mathematical model for multiple linear regression is:

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_6 x_6 + e \tag{1}$$

In the above formula, β_0 is a constant term, and e is a residual term.

Table 3. Regression Coefficients.

	Estimate	Std. Error	T-statistic	P - value	VIF
Constant	13261.4	115.4	114.898	$< 2\times10^{-16}$	-
z_1	3609.3	116.3	31.029	$< 2\times10^{-16}$	1.015
z_3	1967.2	117.7	16.711	$< 2\times10^{-16}$	1.040
z_4	611.9	115.5	5.296	$< 1.27\times10^{-7}$	1.002
z_5	9645.8	115.5	83.522	$< 2\times10^{-16}$	1.001
z_6	-354.3	117.0	-3.029	0.00247	1.026

The model's coefficient of determination R^2 is 0.7504, and the adjusted R^2 is 0.7499, indicating that the model fits quite well. The p-value corresponding to the F-statistic is 2.2×10^{-16}, which is close to 0, suggesting that the joint effect of all independent variables on the dependent variable in the model is significant.

Based on Table 3, it can be noticed that the z_2 is deleted when screening the variables using stepwise regression. The independent effects of each independent variable on the dependent variable are all significant, as shown by the p-values of the t-statistics for the remaining variables being less than 0.05. Therefore, the regression equation can be written as:

$$\hat{Y} = 3261.4 + 3609.3z_1 + 1967.2z_3 + 611.9z_4 + 9645.8z_5 - 354.3z_6 \quad (2)$$

According to Table 3, each variable's variance inflation factor (VIF) is less than 5, which implies there is either minimal or no multicollinearity among the independent variables.

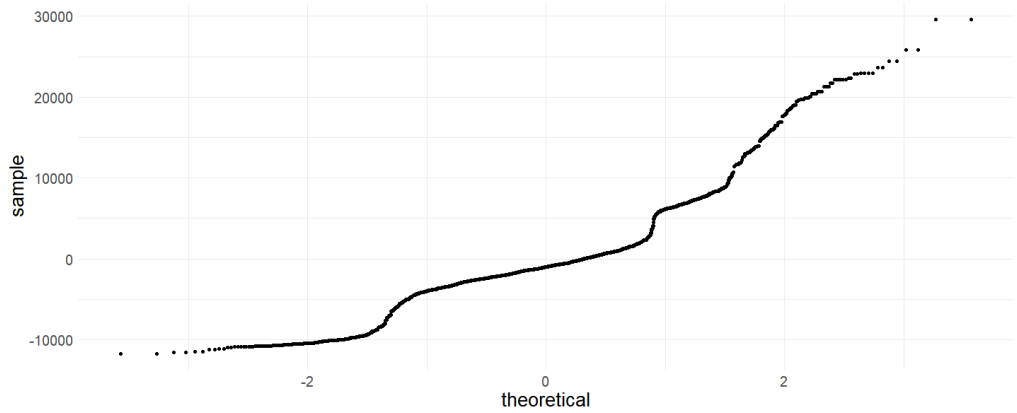

Figure 3. Quantile - Quantile Plot.

The points on Figure 3 roughly form a straight line, meaning that the residuals obey the assumption of normality.

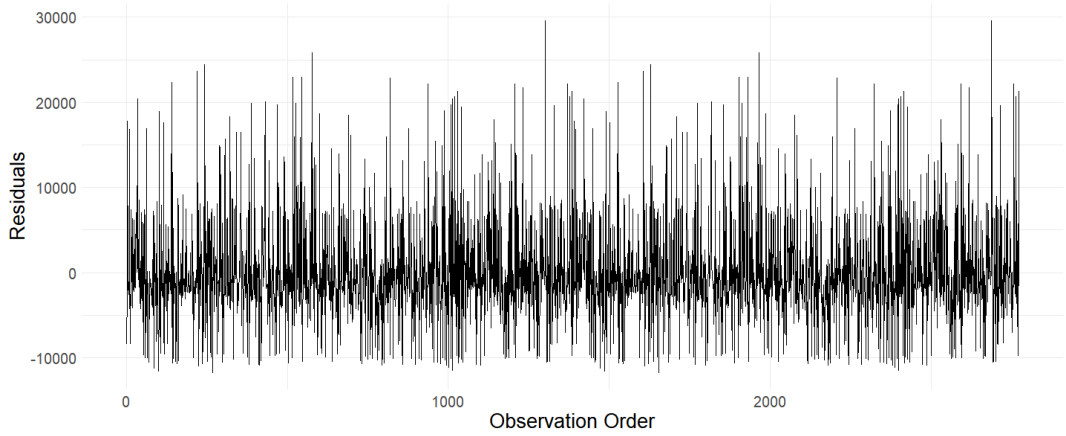

Figure 4. Residuals vs. Order.

As can be seen in Figure 4, all points are randomly distributed around the x-axis with no obvious upward or downward trend, which verifies the zero-mean assumption of the residuals. And all points maintain a relatively stable degree of dispersion throughout the observed series, which illustrates the homoskedasticity assumption of the residuals.

3.4. Discussion

Of all the variables, Age, BMI, and Smoker have the largest regression coefficients, indicating the greatest degree of influence on the dependence. This demonstrates that it is physiological factors that have the greatest impact on medical insurance cost, and goes some way to show that keeping body in good shape and having a good routine can reduce the amount of money spending on medical insurance.

During the screening of variables using stepwise regression, the other variables are retained and highly significant, as expected, but it is surprising that the variable Sex is removed, explaining the insignificant contribution of the change in the dependent variable. This could be because over half of the participants are adolescents, which increases the likelihood that they will have a spouse and, thus, that their medical insurance cost may be related to their spouse. Furthermore, a small number of non-parent participants are minors, suggesting that their medical insurance cost may be connected to their parents. The above phenomenon suggests that the effect of gender on medical insurance expenditures studied in this research may be blurred by household factors.

4. Conclusion

In the descriptive statistical analysis section, this study graphically presents the distributional characteristics of the variables and finds that these characteristics are consistent with the general perception. In the multivariate statistical analysis section, this author finds that five factors, age, BMI, number of children, smoker or non-smoker, and region, have a significant effect on medical insurance spending, with the first four variables having a positive effect and the last variable having a negative effect. The strongest regression coefficients are found in age, BMI, and smoker or non-smoker, indicating that physiological parameters are the main determinants of medical insurance costs. Therefore, it is recommended to stay healthy, especially smoking less, to reduce medical insurance expenses and save money. The model fit is also good. Through correlation analysis, the correlation coefficients between the independent variables are found to be very low, which is then proved using VIF after multivariate statistical analysis. The final residual analysis shows that the residuals obey the normality assumption, the zero-mean assumption and the homoscedasticity assumption.

This study explores the effects of several important variables on medical insurance expenditures through multiple linear regression and summarizes the regression equation. For people who purchase insurance, it predicts costs based on personal background, highlighting the role of health protection in reducing costs. For insurance companies, it aids in precise pricing and personalized products. For policymakers, it clarifies market rules and provides a scientific basis for policy formulation. However, there are still some deficiencies, such as not controlling for the same family background when using the variable Sex. An improved approach would be to collect data on spending on medical insurance for unmarried and married adults separately for the study, exploring the effect of gender on this spending. In addition, more variables should be selected to address the problem that the study has limitations.

References

[1] P. B. Trescott, History of the US Health Care Industry. Salem Press Encyclopedia, (2024).

[2] S. Emon, M. R. Hossain, S. M. M. Hasan, et al. Prediction of Medical Insurance Costs: A SHAP-Enhanced Predictive Analysis for Transparency and Interpretability. 2023 26th International Conference on Computer and Information Technology (ICCIT), Computer and Information Technology (ICCIT), 2023 26th International Conference On, (2023) 1–6.

[3] K. Li. Analysis of Regional Differences in Expenditures of the Health Insurance Fund - Based on the Dagum Gini Coefficient Method. Public Finance Research, 2 (2021) 62-71.

[4] T. Mathur, U. K. Paul, H. N. Prasad, et al. Understanding perception and factors influencing private voluntary health insurance policy subscription in the Lucknow region. International Journal of Health Policy & Management, 4(2) (2015) 75–83.

[5] E. M. Selamat, S. R. Abd Ghani, N. Fitra, et al. Systematic Review of Factors Influencing the Demand for Medical and Health Insurance in Malaysia. International Journal of Public Health Research, 10(2) (2020) 1242–1250.

[6] S. M. Sanjaya, T. S. Zen. Aspects Influencing Personal Life and Health Insurance Purchase. Journal Return, 2(8) (2023) 821–831.

[7] D. Zheng. Research on the Influencing Factors of China's Health Insurance Fund Expenditure (Master's Thesis, Shanghai University of Finance and Economics), (2023).

[8] P. Zhang. A study of the impact of environmental pollution on health insurance expenditures and its mechanisms. Modern Economic Research, 10 (2019) 28-37.

[9] Q. Wang, J. Wang, F. Gao. Who is more important, parents or children? Economic and environmental factors and health insurance purchase. North American Journal of Economics and Finance, 2021 58.

[10] L. Zhou, L. Zhao. A study of the impact of WeChat's exercise health factor on health insurance rates. Science Technology and Industry, 7 (2022) 253-258.