Dean&Francis

ISSN 2959-6157

The Influencing Factors of Healthcare Insurance Coverage Based on the Multiple Linear Regression Model

Jieying Wang

Department of Operations and Risk Management, Lingnan University, Hong Kong, 999077, China
jieyingwang@ln.hk

Abstract:

The rapid growth of the healthcare insurance market in recent years has caused a number of concerns about the financial burden on insured individuals in purchasing health insurance coverage, as well as the sustainability of insurance funds in the future. To address these issues, multiple linear regression is defined as a method to interpret the change in healthcare coverage amount as a result of demographic factors, disease type, and healthcare expenses. Due to the time effects on the original healthcare coverage amount, the lagged healthcare coverage amount has been added. The model is based on a dataset containing 736 patients' records. This relationship is formulated as an equation. The assumptions and multilinear regression analysis - normality, linearity, and extreme values - are examined. The results suggest that gender, body mass index, race, healthcare expenses, and lagged healthcare coverage amount have a significant impact on the level of healthcare coverage amount.

Keywords: MLR, Healthcare insurance coverage, Lag

1. Introduction

Health insurance covers the cost of medical expenses associated with accidents or illnesses [1]. The United States is expected to spend $7.7 trillion on healthcare in 2025, surpassing $5 trillion in 2024, as the amount spent on healthcare in the United States will increase rapidly [2]. The global healthcare insurance market is expected to experience a 5% increase from 2025 to 2033, based on US$2,332.1 billion in 2024 [3]. There are a number of factors that have contributed to the increase in medical insurance costs, including demographic changes, deteriorating health, and rising costs for Medicare and Medicaid programs [4]. Healthcare is becoming increasingly important for both individuals and society at large. It is imperative that the healthcare insurance coverage amount be increased in order to resolve the problem of patients excessively relying on out-of-pocket payments resulting from resource constraints due to the limited number of healthcare resources and poor efficiency of resource allocation [5].

The amount of healthcare coverage is affected by a number of factors, including demographics, socio-economic status, and diseases and health conditions. A majority of previous studies have focused on individual factors, such as social demographic characteristics, socioeconomic characteristics, or specific age groups or populations with particular diseases [6]. In the aftermath of numerous research studies examining the impact of different factors on the amount of healthcare coverage for individuals, some progress has been made in various areas. Firstly, HE is significantly associated with a number of factors, including a person's age and gender [7]. The cost of healthcare increases as people age and becomes the most significant financial burden after the age of 85 [7]. According to research, there is a significant positive correlation between age and healthcare utilization, as well as a much higher rate of health service utilization among the elderly than among the younger population [6]. Secondly, factors such as race and BMI can affect HE to some extent. Race is associated with substantial disparities in health status among individuals [8]. Obesity and overweight (BMI 25 kg/m2) cause an increase of over 150% in healthcare costs for circulatory, respiratory, analgesics, the central nervous system, and malignant diseases [3]. Further, HE varies according to DT. For instance, chronic diseases require long-term treatment and management, resulting in high healthcare costs and claims, while acute diseases require high short-term costs [4]. In the COVID-19 period, Bundorf and Gupta (n.d.) examined the relationship between respondents' race and health insurance coverage [8]. While white populations have stable insurance coverage, other races face obstacles when it comes to obtaining health insurance [8].

In summary, while prior studies have explored healthcare expenditure and insurance coverage to some extent, there remains a gap in understanding the quantitative influence of multiple independent variables on healthcare coverage amounts. This study addresses that gap by providing a comprehensive model to estimate healthcare coverage levels. This study aims to provide new insights into the estimation of healthcare coverage, which can be employed to prepare reserve funds and calculate premiums based on this coverage estimation.

2. Methodology

2.1 Data Source

This paper utilizes data from Synthea, including four datasets: allergies, care plans, observations, and patients [9]. After reintegrating these four datasets, observations with missing data were discarded, leaving 736 observations for analysis. With the aid of generative artificial intelligence technology, Synthea generates synthetic medical claims data that is similar to actual patient data, so as to avoid not only the risk of exposing patients' personal information but also security and privacy issues.

2.2 Variable Selection

The purpose of this study is to examine the factors influencing healthcare coverage amount (HCA) and analyze their ability to predict HCA. As the research period, the period between October 8, 1912, and April 7, 2025, is used, as well as the information of patients and their healthcare utilization generated after October 8, 1912.

The dataset contains a wide range of independent variables. The dependent variable is the amount of healthcare coverage. The independent variables are patients' age, gender, body mass index (BMI), race, DT, and healthcare expenditures (HE). The majority of patients' healthcare records between the ages of 0 and 20 years old. Females account for more than half of the healthcare records. More than 50% of BMI records fall within the range of 20 to 30 kilograms per square meter. More than 80% of the White healthcare records are included. As for DT, there are 37 diseases classified into 5 categories, while special situations and behavioral management account for 60 percent of the total. In terms of HE, medical support organizations tend to charge more than 0.1 million per patient for HE. Moreover, most patients have more than 0.2 million in healthcare coverage, while a portion do not have such coverage. In this study, females are labeled as 1, while males are labeled as 0. Whites are labeled as 1, and other races are labeled as 2. The diseases are classified as 5 different types, special situations and behavioral management for 1, chronic disease management and nursing for 2, nursing plan and record for 3, medical and rehabilitation-related operations for 4, acute disease and injury for 5.

2.3 Model Selection

In this research, multiple linear regression (MLR) is used to establish the linear relationship between the dependent variable and the independent variables. Considering the time lag in healthcare coverage, a first-order lag processing is applied to healthcare coverage. The equation is as follows:

$$\begin{aligned}Y = \beta_0 + \beta_1\text{*Age} + \beta_2\text{*Gender} + \beta_3\text{*BMI} + \beta_4\text{*Race} \\ + \beta_5\text{*HE} + \beta_6\text{* } LHCA\text{ﬁ} + \epsilon\end{aligned} \tag{1}$$

Where Y indicates HCA; β0 refers to the intercept; coefficients β1, β2, … β6 represent the magnitude of the effect of every independent variable on the healthcare insurance coverage, ϵ is the error term.

3. Results and Discussion

3.1 Model Results and Discussion

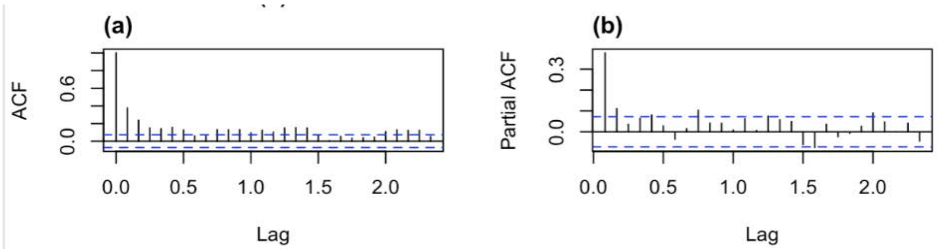

Fig. 1 Autocorrelation and Partial Autocorrelation Functions (Photo/Picture credit: Original).

Fig. 1 presents lagged HCA. Fig. 1 (a) depicts the autocorrelation function, and Fig. (b) shows the partial autocorrelation function. At lag 1 point, there is a significant correlation between the current healthcare coverage amount and the previous period‘s healthcare coverage amount, while the partial autocorrelation coefficient falls within the confidence interval [blue line]. Hence, fitting the autoregressive part of MLR at the order of 1 is appropriate to be applied for weaken the influence of autocorrelation. A statistical analysis of the Lag MLR model and residuals is presented in Table 1. In the Lag MLR model, the F value is 104.143, and the p value is less than 0.001, suggesting that all of the results are statistically significant.

Table 1. Analysis of Variance

	Sum of squares	Df	Mean squared	F value	Significance
Lag MLR	$7.783*10^{13}$	7	$1.112*10^{13}$	104.143	<0.001***
Residuals	$7.783*10^{13}$	729	$1.068*10^{11}$		

A summary of how those independent variables perform in the original MLR model is presented in Table 2. The Durbin-Watson value is 1.240, which is lower than 1.5 and indicates that the residual has a positive autocorrelation. Based on healthcare coverage exhibiting an inertia or lag effect, a new MLR model is generated that incorporates a lag of the amount of healthcare coverage to resolve the positive autocorrelation.

Table 2. Summary of the Original MLR Model

	Regression coefficient	Standard Error	t-value	Pr>|t|	95% Confidence limit VIF Tolerance		Collinearity Diagnostics	
Intercept	-165263.491	80781.312	-2.046	0.041	-323854.534	-6672.448		
Age	-364.173	881.799	-0.413	0.680	-2095.333	1366.988	2.206	0.453
Gender	257017.329	26929.912	9.544	<0.001	204148.135	309886.524	1.044	0.958
BMI	27484.253	2606.803	10.543	<0.001	22366.540	32601.965	1.323	0.756
Race	-146408.660	33595.883	-4.358	<0.001	-212364.585	-80452.735	1.014	0.986
DT	1089.410	15165.463	0.072	0.943	-28683.647	30862.467	2.052	0.487
HE	-0.150	0.018	-8.159	<0.001	-0.187	-0.114	1.788	0.559
R-squared	0.2874							
Adjusted-R	0.2816							
F value	49.14 p<0.001							
D-W value	1.240							

In contrast to Tables 2 and 3, the residual adheres better to the independence assumption, enhancing model reliability. LHCA significantly (p <0.001), suggesting that LHCA possesses independent explanatory power for the original healthcare coverage. The Lag MLR model‘s explanatory power increases when the R-squared values increase. Compared to the original MLR model, the Lag MLR model’s D-W value is 1.872, which is close to 2, as no significant autocorrelation is detected in the residuals, suggesting that the model adequately captures healthcare data.

Fig. 2 depicts the residual plot for the Lag MLR model. There is no apparent linear trend among the points, and the correlation coefficient (0.075) is close to 0, suggesting that there is little or no linear autocorrelation between the residuals lagged by 1 period in this model. Hence, the Lag MLR model is good for fitting residual autocorrelation without exhibiting time-dependent effects.

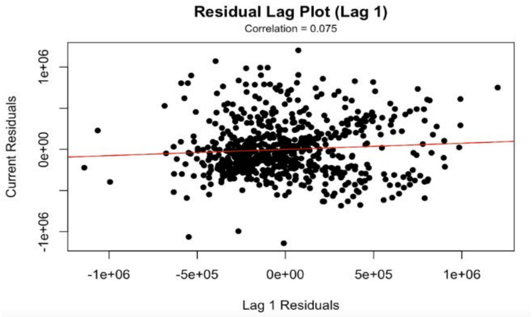

Fig. 2 Residual Lag Plot (Lag 1) (Photo/ Picture credit: Original).

Table 3 contains multiple statistics regarding the independent variables. The R-squared is 0.377, indicating that the model explains 37.7% of the total variation of the independent variable. In this regard, the model has a moderate explanatory power for the dependent variable. Based on the model‘s goodness of fit, the adjusted R is 0.371, which indicates that the addition of the lagged healthcare coverage amount has a positive impact on the model‘s explanatory power. The F-value is 63, and the p-value is less than 0.001, suggesting that the entire multiple regression model is statistically significant. The multiple linear regression equation is

$$\begin{aligned} & HCA = -136267.088 + 494.412 \times \\ & Age + 249251.901 \times Gender + \\ & 21722.062 \times BMI - 142582.887 \times \\ & Race - 12554.529 \times DT - 0.129 \times \\ & HE - 0.293 * LHCA \end{aligned} \tag{2}$$

Furthermore, the multicollinearity diagnostics of the Lag MLR model reveal that every VIF value is lower than 3 (<10), indicating that there is no issue with multicollinearity, and this allows for the standard interpretation of the regression coefficients.

Table 3. Summary of Lag MLR Model

	Regression coefficient	Standard Error	t-value	Pr>|t|	95% Confidence limit VIF		Collinearity Diagnostics	
							Tolerance	
Intercept	-136267.088	74011.079	-1.841	0.066	-281567.706	9033.530		
Age	494.412	810.863	0.610	0.542	-1097.497	2086.322	2.223	0.450
Gender	249251.901	24694.343	10.093	<0.001	200771.277	297732.526	1.045	0.957
BMI	21722.062	2453.225	8.854	<0.001	16905.823	26538.019	1.399	0.715
Race	-142582.887	30753.792	-4.636	<0.001	-202959.101	-82206.183	1.014	0.986
DT	-12554.529	13964.191	-0.899	0.369	-39969.420	14860.362	2.074	0.482
HE	-0.129	-0.017	-7.568	<0.001	-0.162	-0.096	1.819	0.550
LHCA	-0.293	-0.031	9.516	<0.001	-0.232	0.353	1.095	0.913
R-squared	0.377							
Adjusted-R	0.371							
F value	63 p<2.2e-16							
D-W value	1.872							

ISSN 2959-6157

3.2 Model Assumptions and Evaluation

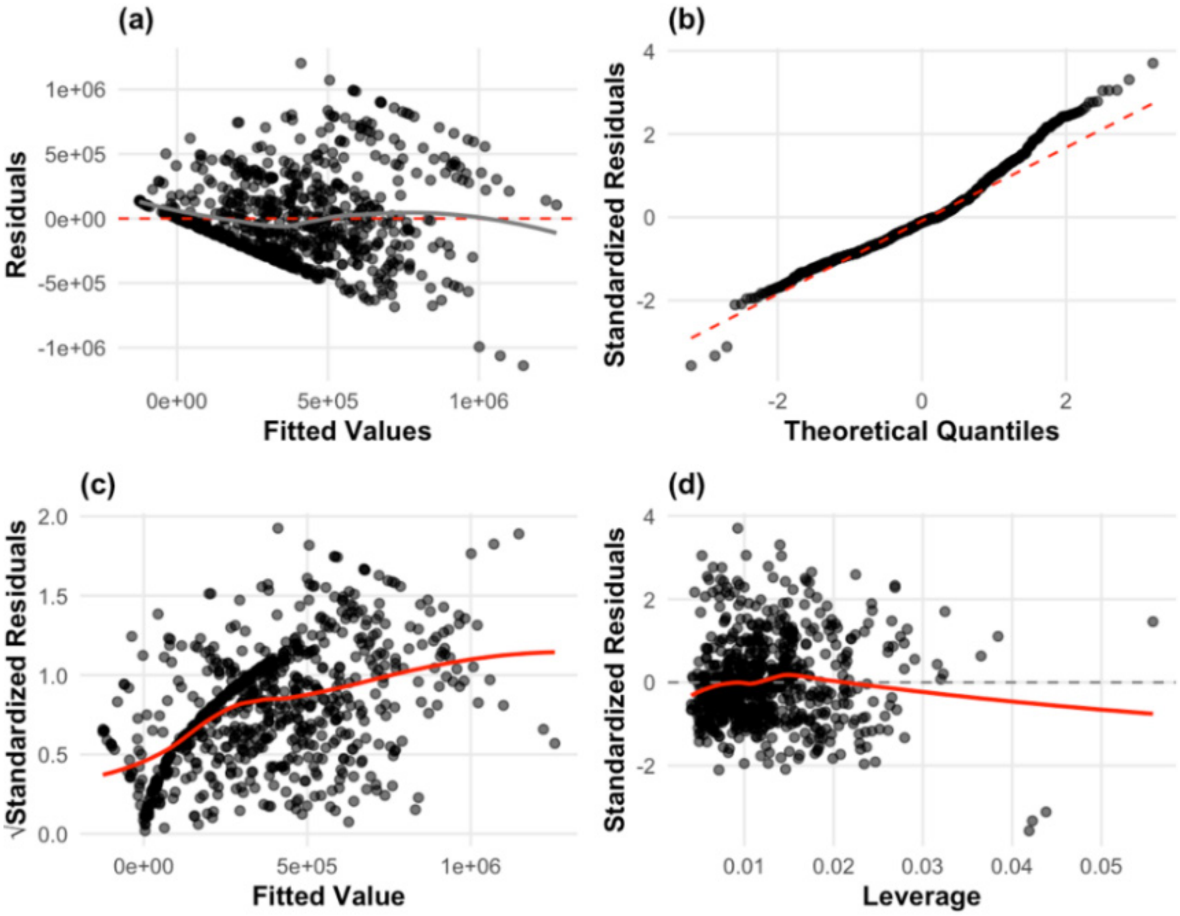

Fig. 3. Lag MLR model diagnostic plot (Photo/Picture credit: Original).

Fig. 3 illustrates the Lag MLR model performance. Fig. 3(a) shows the relationship between fitted values and residuals, while Fig. 3(c) illustrates the relationship between the square root of the normalized residuals and the fitted value. The randomness of residuals indicates that the linearity assumption is not grossly violated, since a linear relationship exists between the predicted value and residuals. In Fig. 3(b), the standardized residuals and quantiles of the theoretical normal distribution are presented. There is a tendency towards normality in the data as a whole. According to Fig. 3(d), the relationship between standardized residuals and leverage value can be seen. The majority of residuals and leverages are within the range of 0 to 0.04.

This research provides a basis for insurance pricing, risk management, reinsurance, and reserve preparation by analyzing influencing factors. Due to a lack of data capacity, model results, and time constraints, some problems were encountered. As the sample size (n=736) is insufficient to represent the entire population, likely, subtle variations in HCA are not detectable. According to the Lag MLR model, 37.73% of the variation of HCA can be explained by this model. This indicates that the model has limited explanatory power. This short-term study processes limited dependent variables LHCA, Lagged MLR model is built rather than building multiple models and comparing them to choose the most optimal model. Kim and Kong examined the relationship between health insurance utilization and demographic, economic, and health status factors by comparing the performances of those independent variables in a stepwise MLR targeting the factor of age [6]. Future development can include capturing seasonal and annual trends, adding economic and health status independent variables, collecting thousands of data across the globe, and implementing non-linear models to account for variation in HCA.

4. Conclusion

As a result of this research, a Lag MLR model is created to study the effect of age, gender, BMI, race, DT and HE on the amount of HCA. In parallel, the amount of healthcare coverage has been increasing year after year. Based on the model and the aforementioned Figs and tables, the five main factors, gender, BMI, race, HE, and LHCA, have a significant impact on the level of HCA. Therefore, healthcare insurance premium pricing could be optimized by refining current risk classifications and implementing differentiated pricing to optimize the healthcare insurance structure and to prepare a sustainable reserve for healthcare insurance claims. The reserve prediction could be optimized through healthcare claim predictions. It is also possible for the reinsurance company to reassess the underwriting risks and the method of reinsurance used by the insurance company. Lack of data, model interpretation, time constraints, and some deviations in the results will affect the accuracy of the results. In future research, this research result may be further developed by capturing more time effects, adding independent variables, and introducing applicable models.

References

[1] Insurance Authority. Functions of medical insurance. 18/1/2024. https://www.ia.org.hk/en/medical_insurance/index.html

[2] CMS. CMS releases 2023-2032 National Health Expenditure Projections. 4/12/2022. https://www.cms.gov/newsroom/press-releases/cms-releases-2023-2032-national-health-expenditure-projections

[3] Health Insurance Market Report, By coverage type (hospitalization, outpatient services, prescription drugs, dental coverage, vision coverage), distribution channel (insurance brokers, direct sales, employer-sponsored plans, government schemes, agents), end-user (individuals, employers, senior citizens, students, expatriates), and Regions 2025-2033. (n.d.).

[4] Fiscella, Kevin MD, MPH; Williams, David R. PhD, MPH. Health disparities based on socioeconomic inequities: Implications for Urban Health Care. Academic Medicine, 2004, 79(12), 1139-1147.

[5] Kodali P. B. Achieving universal health coverage in low- and middle-income countries: challenges for policy post-pandemic and beyond. Risk management and healthcare policy, 2023, 607–621.

[6] Kong, Na Young, and Dong Hee Kim. Factors influencing health care use by health insurance subscribers and medical aid beneficiaries: a study based on data from the Korea welfare panel study database. BMC Public Health, 2020, 20(1): 1133.

[7] Bae CY, Kim BS, Cho KH, Kim IH, Kim JH, Kim JH. 10-year follow-up study on medical expenses and medical care use according to biological age: National Health Insurance Service Health Screening Cohort (NHIS-HealS 2002~2019). PLoS One. 2023.

[8] Bundorf MK, Gupta S, Kim C. Trends in US health insurance coverage during the COVID-19 pandemic. JAMA Health Forum. 2021.

[9] Walonoski J, Hall D, Bates KM, Farris MH, Dagher J, Downs ME, Sivek RT, Wellner B, Gregorowicz A, Hadley M, Campion FX, Levine L, Wacome K, Emmer G, Kemmer A, Malik M, Hughes J, Granger E, Russell S. The “Coherent Data Set”: Combining patient data and imaging in a comprehensive, synthetic health record. Electronics. 2022; 11(8):1199.