Health Insurance Annual Premium Forecast Analysis

Siwen Chen

Bachelor of science, queen mary, london, e1 4ns, united kindom

ah21512@qmul.ac.uk

Abstract. This study focuses on predicting an important socioeconomic indicator: health insurance costs and provides an in-depth exploration of the impact of multiple factors on insurance costs. The data used in the study comes from a representative health insurance company, and its data individuals include indicators of multiple dimensions, such as the age, gender, body fat percentage, family size of the insured, as well as whether they have smoking habits and the specific region where they are located. and other information. In order to accurately reveal the impact of these factors on insurance premiums, we adopted a machine learning model, the Lasso regression model, for modeling and prediction, supplemented by the calculation of correlation coefficients to quantify the strength of the relationship between these factors and insurance premiums. After in-depth exploration and analysis, the research results show that among all factors considered, age, body fat percentage and whether you smoke have a significant impact. It is particularly noteworthy that the factor of smoking has the most significant impact on insurance costs. In addition, the study also revealed that women and insured persons living in the southeast region tend to choose higher premiums. These research results not only have a certain enlightenment effect on theoretical research, but also have significant reference value for the practice of the insurance industry. It can help insurance companies more accurately identify and evaluate potential risks, and set more scientific and reasonable insurance rates accordingly.

Keywords: Health insurance premiums, Lasso regression model, predictive analysis, assessment of influencing factors.

1. Introduction

Health is the eternal theme pursued by human beings. With the progress of society and the development of civilization, humans are paying more and more attention to health. At the same time, the increasingly high medical costs have made people feel unbearable the risk of wealth loss caused by illness or accidental injury. Health insurance is an effective means of spreading and transferring this risk. At present, my country's health insurance market is growing rapidly. Since 2000, my country's commercial health insurance premium income has maintained rapid growth at an average annual growth rate of 1.4 times [1]. By the end of 2006, my country's annual commercial health insurance premiums had reached 37.6 billion yuan [2]. The reason is that on the one hand, people gradually realize the objectivity and universality of the risks of disease and accidental injury, and their care for their own lives and health continues to increase; on the other hand, with the increase in family economic income, people are more interested in health insurance. purchasing power is increasing.

At present, my country's health insurance products are mainly one-year short-term insurance, which has great limitations for policyholders. For example, insurance companies generally stipulate that insurance can only be renewed until the age of 65, and premiums must be paid year by year and must also be underwritten year by year and based on the insured's physical health, financial status, changes in occupational risks and other factors, the customer will be asked to increase the premium or even refuse the insurance when renewing the policy [3]. Long-term or lifelong health insurance allows the insured to not have to worry about increasing premiums or being refused insurance when renewing the policy, which meets people's long-term medical security needs. Long-term health insurance gives a lifelong promise of happiness, so people's demand for long-term health insurance must be will rise day by day. Now our country.

Many life and health insurance companies are also launching some long-term health insurance products. For example, China Life Insurance Co., Ltd. launched "Care for Life Long-term Health Insurance" [4]. Ping An Life Insurance Co., Ltd. of China launched "Guard Life Lifelong Health Insurance" [5] and Chinese People's Health Insurance Co., Ltd. has released my country's first long-term care insurance product "Worry-Free Long-term Care Personal Health Insurance" [6] and so on. It should be said that the spring for the development of my country's long-term health insurance has arrived.

Of course, Chinese scholars have also conducted valuable exploratory research on health insurance actuarial science in recent years. One of the more influential ones is Dr. Chen Tao of Southwestern University of Finance and Economics. His representative work is his doctoral thesis "Medical Insurance Actuarial and Risk Control Methods" (now publishing) [7]. However, this paper only studied medical insurance in health insurance, and mainly focused on the actuarial aspects of short-term medical insurance, and did not involve actuarial content such as critical illness insurance. In addition, Dr. Jing Tao of the University of International Business and Economics has also made certain achievements in the field of long-term care insurance. In her doctoral thesis "Research on Long-term Care Insurance" (published) [8], she innovatively proposed a three-step approach to long-term care insurance in my country from a macro level. development model, but the paper also did not study the actuarial issues of long-term care insurance at the micro level. Judging from the overall research status, my country's research in the field of long-term health insurance actuarial science is still in its infancy.

With the development of social economy, the long-term health insurance market presents huge potential and challenges. In the future development process, we will inevitably encounter various complex problems, among which the actuarial accuracy of long-term health insurance is a key factor in controlling health insurance business risks. To this end, this article chooses long-term health insurance actuarial science as the research topic, aiming to provide accurate predictions of the future health insurance market through in-depth exploration and research, and provide reference for related research. Our research will be based on the customer data of insurance companies, systematically analyze various important indicators that affect premium pricing, and strive to provide new observations and interpretations on premium pricing of long-term health insurance. Further, we will predict customers' annual premiums in order to provide insurance companies with more precise guidance in risk control and capital planning. The importance of this study is self-evident. It not only helps to enrich theoretical research in the field of health insurance, but also provides strong theoretical support for actual insurance operation and management. We hope that this research can provide some reference and inspiration for research in health insurance-related fields. We also hope that it can provide valuable thinking and reference for our future management and sustainable development of long-term health insurance.

2. Data description

This dataset constructs a comprehensive overview of demographic information derived from a health insurance company. In order to accurately study and determine the various factors affecting health insurance costs to further optimize the premium pricing system, the company selected seven relevant indicators for information collection and research. This data set covers various age groups, genders, and regions, including lifestyle habits and physical health status (BMI) and other aspects of information, forming a detailed statistical file. Of these seven indicators, six serve as independent variables that will play an important role in the subsequent analysis, while the last indicator serves as the dependent variable, revealing specific health insurance costs. All data were strictly anonymized to ensure the protection of participant privacy while retaining nutritional value for meaningful data analysis to examine how variables impact health insurance costs, a core goal of this study. Through in-depth mining and research of these data, this study will fully understand the various factors that

affect health insurance costs and provide strong support for insurance companies to develop more accurate and targeted pricing strategies.

Table 1. Indicators and Descriptions for Studying Factors Impacting Health Insurance Costs

Variable	Variable description
Age	The age distribution of the subjects ranged from 18 to 64 years old, showing a roughly even distribution. It can be tentatively hypothesized that age is related to insurance premiums, with older adults likely being associated with higher premiums due to increased health risks.
Gender	This variable was binary, dividing subjects into male (n=676) and female (n=662) groups, which were fairly balanced in size.
Body Mass Index (BMI)	BMI is calculated by dividing weight in kilograms by height in meters squared. It is a reliable indicator of an individual's physical condition. The generally considered standard range is between 18.5 and 24.
Number of dependents	This variable ranges mostly from 0 to 3, with a few exceptions being 4 and 5. Presumably, individuals with more dependents may be associated with lower insurance premiums due to underlying financial constraints.
Smoking status	This binary variable records whether the subject smokes (yes or no). Presumably, smokers may have to pay higher premiums due to the increased health risks.
Area	Classification according to eight geographical locations may be related to the level of economic development of the region. This factor deserves further analysis.
Health insurance costs	Reflecting annual health insurance costs, this outcome variable may be affected by the factors mentioned above. Key to our investigation is the derivation of statistical models or assessments of the relative impact of factors that help predict health insurance costs. This forms the crux of this article.

The data set under study consists of seven columns, with the first six columns representing the independent variables and the last column representing the dependent variable, the outcome. Our study aims to provide empirical insights into the determinants of health insurance premiums by studying the above-mentioned variables and their complex interrelationships.

3. Method

3.1. Data preprocessing

Chapter 1 has already described the source of the data and made some simple descriptions of the data. In the mathematical model that will be used, simple preprocessing of the data is required. For categorical data such as "male" and "female" for gender, and "yes" and "no" for smokers, we digitize the different categories by encoding them with 0 and 1; The data is normalized to facilitate subsequent calculations. The specific normalization formula is:

As for the final prediction target, premium data, the original value is still maintained for later prediction.

$$x_i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (1)$$

Where X represents the variable data to be normalized, it can be regarded as a set of vectors, indicating the ith value in the vector, indicating the ith value after normalization.

As for the final prediction target, premium data, the original value is still maintained for later prediction.

3.2. Lasso regression model

The regression model is a method often used in processing data. Conventional linear regression model is generally used when there are only one or fewer independent variables. When the prediction result is affected by multiple variables, we often use more advanced algorithms. This article uses the LASSO (Least absolute shrinkage and selection operator) algorithm [9]. The Lasso model is a regression method suitable for multicollinearity problems and can achieve variable selection while estimating parameters.

In the nascent stages of constructing a model, a liberal inclusion of independent variables is typically adopted. This approach helps to mitigate the potential bias in the model induced by the absence of crucial independent elements. However, the process of model formation necessitates indentifying a collection of independent variables that possess the most potent explanatory capacity for the dependent variable. This implies enhancing the model's elucidation faculty and its ability to predict with precision through the selection of independent variables (choices of indicators or fields). The selection of the most effective indicators is a tremendously cardinal aspect in the procedure of statistical model formation. Lasso, a technique of estimation, has the ability to condense the set of indicators.

The Lasso technique is recognized as a method of compressed estimation [10]. By forming a penalty function, this approach generates a more sophisticated model, forcibly reducing some coefficients to zero and compressing others. Owing to this, it maintains the subset shrinkage's advantage and is utilized for biased estimation when handling intricate collinearity within datasets. A distinct feature of Lasso regression is its ability to perform variable selection and complexity regulation while fitting into a generalized linear model. Consequently, regardless of the nature of the targeted dependent variable - whether it's continuous, binary, or multivariate discrete - employing Lasso regression modeling is feasible for predictions. Variable selection in this context implies not incorporating all variables within the model fitting but rather selectively inserting variables in order to achieve superior performance parameters. Complexity regulation pertains to the control of model complexity through a variety of parameters to prevent overfitting.

With respect to models of a linear nature, there exists a direct correlation between the model's complexity and the quantity of variables it encompasses. An augmentation in the variable count escalates the complexity level of the implicated model. Oftentimes, an inflated variable count may yield a model that appears superior upon initial fitting, yet it may pose an imminent risk of succumbing to an overfitting predicament [11]. In such an event, when the model is subjected to validation against a new data set, the outcome tends to disappoint. As a general rule of thumb, circumstances where the variable count significantly overshadows the data point tally, or when a singular variable encompasses an oversized repository of unique values, there exists a likelihood for overfitting to occur.

The intricacy alteration of LASSO regression is steered by the parameter symbolized as λ [12]. An increment in λ escalates the penalty for linear blueprints with numerous variables, eventually fabricating a model with a reduction in variables. LASSO regression alongside Ridge regression is attributed to a cohort of generalized linear blueprints denominated as Elastic Net. In unison with the parameter λ displaying similar effects, this model family is also regulated by an additional parameter α, which governs the model's activity whilst handling data with high correlations. Within this specified family, the model for LASSO regression stipulates $\alpha=1$, the Ridge regression model conditions $\alpha=0$, whereas the α value boundary for other customary Elastic Net models is $0<\alpha<1$.

At present, Stanford statistician Trevor Hastie, creator of LASSO regression, proffers the optimal R package for applying generalized linear models -- 'glmnet' [13]. The distinguishing feature embodied by this package is its capacity to accommodate a myriad of distinct λ values. Each subsequent fitting process harnesses the outcome of its predecessor, thereby drastically enhancing the computational efficiency. Incorporated in its offerings is the provision of parallel computation capability, enabling either a single computer's multiple cores, or a network of computers to be marshalled, thereby curtailing the calculation duration significantly.

Lasso regression is a compressed estimation method that replaces the least squares method. The basic idea of Lasso is to build an L1 regularized model [14]. During the model-building process, some coefficients will be compressed and some coefficients will be set to 0. When the model training is completed, these parameters with weights equal to 0 can be discarded, thereby making the model It is simpler and effectively preventing model overfitting. It is widely used for fitting and variable selection of multicollinear data.

3.3. Regression algorithm

For ordinary linear models, when X is a column-full design matrix:

$$Y = \beta X + \varepsilon \tag{2}$$

In the formula, X represents the independent variable, Y represents the dependent variable, and represents the coefficient and intercept value of linear fitting respectively.

The regression coefficient can be obtained by the least squares algorithm:

$$\hat{\beta} = \arg min||Y - X\beta||^2 = (X^T X)^{-1} X^T Y$$

When X is not a column-full design matrix, the ordinary least squares method is no longer suitable for solving the regression coefficients. At this time, the penalty method is introduced, which takes the value of the minimum penalty likelihood function as the estimated value of the regression coefficient:

$$\hat{\beta} = \arg min||Y - X\beta||^2 + P_\lambda(|\beta|) \tag{3}$$

The penalty item is, and its specific expression is:

$$P_\lambda(|\beta|) = \lambda \sum_{j=1}^{d} |\beta_j|^m \tag{4}$$

Where m≥0, λ is the adjustment parameter. Represents the fitting coefficient corresponding to multiple independent variables. Let m=1 here to obtain the Lasso penalty term.

4. Results

4.1. Data results

After data preprocessing, the current representation of the data is as follows (take the first few rows of the data as examples):

Table 2. Personal health data and insurance cost analysis

No.	Age	Bmi	Children	Sex	Smoker	Charges	Northeast	Northwest	Southeast
1	0.02	0.32	0	0	1	16884.92	0	0	0
2	0.00	0.48	0.2	1	0	1725.55	0	0	0
3	0.22	0.46	0.6	1	0	4449.46	0	0	0

Call the glmnet package in R language, use age, gender, number of children, marital status, smoking status, and area of residence as independent variables, use premium data as dependent variables, set the random seed value to 2023, and the result is as follows:

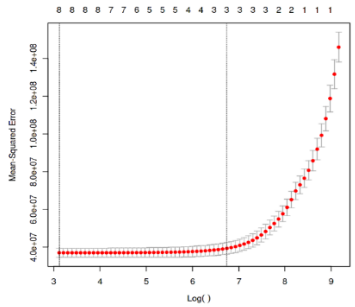

Figure 1. Analysis of λ value and coefficient matrix corresponding to the minimum mean square error

As shown in the figure, the abscissa data of the red point corresponding to the minimum mean square error is 3.115. Since the abscissa uses natural logarithmic coordinates, the corresponding λ value is e^3.115≈22.534. According to the selected λ value, the corresponding the coefficient matrix result is:

Table 3. Analysis of the impact of factors on insurance rates

Age	Bmi	Children	Sex	Smoker	Northeast	Northwest	Southeast
11757	12404	2289	-80	23783.27	873.13	520.85	-60.28

(1) In the above table, the magnitudes of age, body fat percentage and whether you smoke are all above 10^5, indicating that these three data have a greater impact on premiums.

(2) In the above table, the absolute value of the data on whether you smoke is the largest, indicating that smokers have the greatest impact on premiums and can most influence customers' decisions to choose premiums.

(3) In the above table, the values for gender and southwest region are negative. According to the coding logic, the gender is 1 for men and 0 for women, indicating that compared with men, women's willingness to pay higher premiums has slightly increased; similarly In the southeast region, the willingness to pay higher premiums has slightly declined.

According to the above fitting coefficient, calculate the actual mean square error and the mean square error of the prediction result, and obtain the correlation coefficient based on their values. The specific formula is:

$$SST = \sum (y_i - avg(Y))^2$$

1) Total Sum of Squares

$$SSE = \sum (y_i' - avg(Y))^2$$

2) Sum of Squared Errors

$$R^2 = 1 - SSE/SST$$

3) Coefficient of Determination

In the above formulas, SST represents the mean square error calculated based on the actual data, SSE represents the mean square error of the predicted data calculated based on the above fitting coefficient, the correlation coefficient is expressed as 1 and the actual mean square error and the predicted mean square The difference in the error ratio. Generally speaking, the closer the value is to 1, the better the fitting effect. The correlation coefficient of the fitted data this time can be considered to have a better fit to the data.

5. Conclusion

Upon examining the contributory factors impacting health insurance costs, research has elucidated that age, Body Mass Index (BMI), and smoking status are primary determinants. Of these contributory factors, smoking status exerts the most significant influence, thereby requiring smokers to bear higher insurance premiums. Moreover, there is a slight propensity among women to opt for higher insurance premiums as compared to men.

In terms of regional differences, residents in the southeast region demonstrate less inclination to pay higher premiums relative to those in other regions. These empirical findings underscore the necessity for health insurance companies to judiciously consider variables such as the policyholder's age, health status indicated by BMI, and lifestyle habits such as smoking. These parameters allow for risk to be more accurately gauged and insurance rates to be set effectively.

Furthermore, gender and geographical location also necessitate consideration in order to accommodate the varied needs of different demographic groups. As prospective policyholders, individuals can leverage this information to better understand and predict changes in health insurance costs, thereby enabling more informed decision-making.

It's worth noting that these findings are not static. They evolve annually, necessitating that insurance providers monitor market trends closely and intermittently update their premium forecast models. This ensures that the accuracy of forecasts is maintained and that their insurance products remain competitive.

References

[1] Hartman, M., Martin, A. B., Benson, J., Catlin, A., & National Health Expenditure Accounts Team. (2020). National Health Care Spending In 2018: Growth Driven By Accelerations In Medicare And Private Insurance Spending: US health care spending increased 4.6 percent to reach $3.6 trillion in 2018, a faster growth rate than that of 4.2 percent in 2017 but the same rate as in 2016. Health Affairs, 39 (1), 8-17.

[2] Cheng, T. M. (2008). China's latest health reforms: a conversation with Chinese Health Minister Chen Zhu. Health Affairs, 27 (4), 1103-1110.

[3] Stone, D. A. (1993). The struggle for the soul of health insurance. Journal of Health Politics, Policy and Law, 18 (2), 287-317.

[4] Chen, L., Zhang, X., & Xu, X. (2020). Health insurance and long-term care services for the disabled elderly in China: based on CHARLS data. Risk management and healthcare policy, 155-162.

[5] AlZahrani, A. M., BinDajam, O. S., AlGhamdi, S. A., AlQarni, S. S., & Farahat, F. M. (2022). Quality of care provided to diabetic patients attending primary health care centers in National Guard in Makkah Region, Saudi Arabia. Journal of Family Medicine and Primary Care, 11 (6), 2900.

[6] Gauvin, F. P., Wilson, M. G., & Lavis, J. N. (2017). Evidence Brief: Taking a Step Towards Achieving Worry-free Surgery in Ontario.

[7] Chen Tao. (2002). Medical insurance actuarial and risk control methods. Southwestern Finance and Economics Publishing House.

[8] Jing Tao. (2005). Research on long-term care insurance. University of International Business and Economics, PhD thesis.

[9] Januaviani, T. M. A., & Bon, A. T. (2019, March). The LASSO (Least absolute shrinkage and selection operator) method to predict indonesian foreign exchange deposit data. In Proceedings of the International Conference on Industrial Engineering and Operations Management (pp. 5-7).

[10] Maillard, O., & Munos, R. (2009). Compressed least-squares regression. Advances in neural information processing systems, 22.

[11] Gasparrini, A. (2011). Distributed lag linear and non-linear models in R: the package dlnm. Journal of statistical software, 43 (8), 1.

[12] Maleki, A., Anitori, L., Yang, Z., & Baraniuk, R. G. (2013). Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). IEEE Transactions on Information Theory, 59 (7), 4290-4308.

[13] Efron, B., & Hastie, T. (2021). Computer age statistical inference, student edition: algorithms, evidence, and data science (Vol. 6). Cambridge University Press.

[14] Krämer, N., Schäfer, J., & Boulesteix, A. L. (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC bioinformatics, 10, 1-24.