An Empirical Study on the Factors Affecting Life Insurance Charge Premiums

Jingzhi Zhang*

*Corresponding author. Email: 2218277049@qq.com

Department of Western Michigan, Guizhou University of Finance and Economy, 550025, China

Abstract. Predicting an individual's annual insurance premium can help insurance companies evaluate risk and pricing more accurately, thus calculating insurance costs more effectively. With this valuable information, analysts can have a better understanding of consumers' spending habits and the impact on their premium payments. In this paper, the linear regression model is used to construct the regression equation between the influencing factors to find the correlation between each factor and the insurance charge. Moreover, through factor analysis, the most important factor named 'Physique health factors' affecting insurance charges are found. Finally, through comparative analysis, the most critical influencing factors for large insurance charges and small insurance charges are found, which are the presence or absence of a bad appetite factor and Physical health factors. Understanding the patterns in this analysis is crucial for insurance companies as they need to ensure that their insurance fees cover potential claims and other expenses while also staying competitive and attracting more customers. Additionally, predicting an individual's annual insurance premium can help individuals understand their insurance needs and budget, making more informed decisions.

Keywords: Insurance charge, Multiple linear regression model, Factor Analysis, Comparative Analysis

1 Introduction

Insurance companies need to assess the risks and needs of each customer to provide them with the right insurance products and price them. Therefore, by analyzing the factors that affect the cost of life insurance, it is possible to predict the insurance cost of each customer and formulate insurance strategies and pricing schemes accordingly. The core of life insurance is the customer. To maintain a high level of profitability, every enterprise is trying to attract new and retain old customers. This helps insurers maintain strong competitiveness in competition with their competitors. Only by understanding the needs and expectations of customers can we provide customers with better products and services [1]. This paper can provide beneficial knowledge on how age, gender, and lifestyle choices can influence an individual's insurance costs. Such knowledge can be highly useful in creating insurance policies or developing marketing strategies aimed at specific demographics.

From the age of the home to where you live, there are many things that providers look at when setting property insurance charges. There are many factors at work. At this stage, it is already possible to predict a person's annual insurance cost by analyzing aspects of people's daily home life, such as rebuild or replacement costs, home location, and pets [2]. In addition, Machine

ICBBEM 2024, March 29-31, Wuhan, People's Republic of China

DOI 10.4108/eai.29-3-2024.2347316

learning and artificial intelligence techniques are already widely used to predict an individual's insurance premium. Statistical model-based approach: This method uses historical data and statistical models to predict future insurance costs. Commonly used statistical models include linear regression, logistic regression, and decision trees. Machine learning-based approach: This method uses machine learning algorithms to discover patterns and patterns in data and predict future insurance premiums. Commonly used machine learning algorithms include neural networks, support vector machines, and random forests [3].

In this work, this paper combines customer behavior factors, such as whether customers smoke, and basic customer information factors, such as age, gender, etc., by constructing a linear regression model to find the influence of customer behavior factors and customer basic information factors on how much a person pays insurance premiums a year. Then, through factor analysis, the customer's behavioral factors become the most important main factors, affecting the payment of insurance premiums. Finally, through comparative analysis, it is found that the behavioral factors of a customer will greatly affect the change in insurance costs. It can be inferred that the amount of insurance costs today is greatly influenced by customer behavior factors.

2 Related work

Kaur and Negi, the researchers employed factor analysis and discovered that customer Fulfillment from the context of insurance policies hinges on several factors including customized and timely service, brand USP, considerate employees, and price immunity. The study revealed that product offerings are the most significant factor influencing overall customer satisfaction, even if the after-sale service is not entirely satisfactory. They even found that among people who live longer, men live longer than women, while insurers are equally satisfied with this in the public and private sectors [4].

Manuel surveyed to investigate how customers view the life insurance policy in the city of Kottayam. An exploratory research method is adopted in this paper. The survey was aimed only at consumers in the city of Kottayam. Fifty respondents of all ages were selected. This paper mainly discusses the main influencing factors on life insurance customers' purchase of insurance, including investment income, corporation esteem, service standard of insurance companies, and standards of quality for products and services insurance companies. Among the respondents, there is the largest number of people aged between 19 and 28, mainly men, 5001 to 10000, mainly in the private sector [5].

Singh et al.; To explore the service quality of life insurance cooperations and the impact of life insurance companies on customer satisfaction, a questionnaire survey was conducted in the central New Territories area of Delhi. They surveyed 139 respondents and found a quantity of factors, including response and security factors, convenience factors, visible factors, and empathy factors. The researchers also found that only among the respondents, age had a significant effect on their choice. In terms of demographic characteristics, gender, educational background, and annual income have little impact on individual insurance behavior [6].

Ganesh Dash; The purpose of the study was to identify the demographic and socioeconomic attributes of prospective customers that impact their decision to purchase a life insurance policy.

The variables considered in the study included the customer's residential location, the insurance company from which the policy was purchased, and the annual premium amount (price). The study was conducted in rural Odisha, with a sample of over 400 individuals who had purchased life insurance policies. [7]

Sindhuja R and Dr.M.P. Kumaran conducted exploration in Coimbatore city, Tamil Nadu, and sample size of 110 respondents to analyze the preferences of customers while life policy investment decision-making. The study aims to examine various aspects of life insurance policy purchase decisions among respondents, including the preferred policy type, preferred insurance providers, key features that influenced the purchase decision, specific benefits sought, satisfaction levels with the purchased policy, and challenges faced during the decision-making process. Additionally, the study investigates the correlation between age and preferred policy type [8].

3 Methodological Framework

3.1 Description of the data set

The research methodology entails gathering information from A Study of Customers Insurance Charges by Bob Wakefield. The dataset comprises 1337 applications that possess 5 attributes, which depict the traits of customers seeking life insurance. The dataset encompasses comprehensive information regarding insurance customers, encompassing their age, gender, BMI, number of dependents, smoking habits, and region. This data has been collected from multiple sources for each individual customer [9,10]. In the following analysis, due to the use of SPSS with the help of research and analysis, so as to facilitate the analysis, the sex string is converted to sexcode numbers, that is, 1 represents female and 2 represents male; The smoker string is converted to smokercode numbers, that is, 1 represents the smoker, and 2 represents the non-smoker smokers; The region string is converted to a regioncode number, i.e. 1 for Southwest, 2 for Southeast, 3 for Northwest, and 4 for Northeast.

3.2 Multiple linear regression model

In this study, the outcome variable (the dependent variable) was the charges of insurance, and the independent variables were the suspected influencing factors of the premium, i.e., whether they smoked, age, BMI, number of children owned, region, and gender. Due to the use of the step-by-step method, the independent variables of region and gender have been excluded and have no significant effect on insurance costs.

Table 1. Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. The error in the Estimate	R Square Change	Durbin Watson
1	.866[d]	.750	.749	6067.787249	.002	2.087

Note: d. Predictors: (Constant), smokercode, age, bmi, children

e. Dependent Variable: charges

Table 2. ANOVAa

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.470E+11	4	3.675E+10	998.123	.000[e]
	Residual	4.908E+11	1333	368042.10		
	Total	1.961E+11	1337			

Note: e. Predictors: (Constant), smokercode, age, BMI, children

Table 3. Coefficientsa

Model	Unstandardized B	Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	35520.030		29.846	<.001		
	smokercode	-23811.400	-.794	-57.904	.000	.999	1.001
	age	257.850	.299	21.675	<.001	.986	1.014
	BMI	321.851	.162	11.756	<.001	.988	1.012
	children	473.502	.047	3.436	<.001	.988	1.002

Note: a. Dependent Variable: charges

Table 4. Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance	Collinearity Statistics VIF	Collinearity Statistics Minimum Tolerance
1	sexcode	-.005[e]	-.386	.700	-.011	.991	1.009	.985
	regioncode	-.005[e]	-.375	.708	-.010	.969	1.032	.961

Note: a. Dependent Variable: charges. e. Predictors: (Constant), smokercode, age, BMI, children

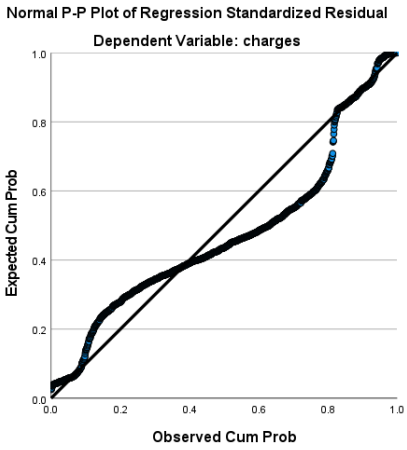

Fig. 1. Normal P-P Plot of Regression Standardized Residual

Analysis: The analysis of the above results is as follows:

1. Table 1 demonstrates that the regression model fits well with an R^2 value of .749, indicating that the independent variable can account for 74.9% of the variance in the dependent variable. This suggests that 74.9% of the variations in insurance charges can be attributed to factors such as smoking status, age, BMI, and number of children.

2. As Tables 2 and 3 show the independent variables of significance $P<0.05$ were whether smoking or not, age, BMI, and number of children had four independent variables, which could significantly affect the change in insurance premium. Furthermore, a negative regression coefficient for smoking means that this independent variable significantly negatively influences the change in insurance charges, i.e., the insurance charges are significantly lower for non-smokers. The positive nature of the regression coefficients of age, BMI, and the number of children implies that the three independent variables have a significant positive impact on the fluctuations in insurance charges, that is, the younger the population, the smaller the BMI index, and the fewer the number of children, the premium is significantly lower.

3. Based on the above analysis, the quantitative relationship (regression equation) between whether or not to smoke, age, BMI, and the number of children is obtained as follows:

Insurance charge = -23811.400 * whether you smoke or not + 257.850 * age + 321.851 * BMI + 473.502 * number of children with + 35520.030

Moreover, it should be noted that the analysis presented here is solely focused on the outcomes of the regression analysis. As for how accurate this conclusion is and how credible it is, it is necessary to judge the regression model this time:

Diagnostic 1: The linear regression model necessitates the absence of multicollinearity among the independent variables. Table 4 indicates that none of the independent variables have a VIF value greater than 5, indicating the absence of multiple collinearities among the variables, and that diagnosis 1 has been successfully validated.

Diagnosis 2: The normality assumption holds that the residuals of the linear regression model should be normally distributed.

Diagnostic 2 has been satisfied as the P-P plot displays diagonal scatters, indicating that the residuals of this regression model conform to a normal distribution. (See Figure 1).

Diagnosis 3: The linear regression model assumes that there is no autocorrelation between observations.

As table 1 shows the statistic that examines whether there is a sequence correlation between the sample data is DW (Durbin-Watson), and the DW value is around 2, which means that there is no sequence correlation between the sample data.

3.3 Factor Analysis

Through the study of the influencing factors, it is found that the commonness of the influencing factors, namely factors. The factors were developed to simplify the overall complexity of the data, utilizing methods such as extraction of the sum of squared loads, scree plots, and the rotated component matrix in order to identify underlying interdependencies.

However, on account of the factor loading being very light, one variable is removed, and that is age. Although the ability of BMI to explain insurance charge = 63.1% < 90% in the Communalities is not strong enough, the extraction of age, number of children, and BMI in the Total Variance Explanation in the Table 7 can cumulatively explain 77.939% of this set of data, that is, the interpretation of insurance charge is very good. And it can be seen that the factor of the number of children and BMI factor are approximately on the same level through the Scree Plot (see Figure 2). Therefore, three factors are extracted for analysis in the following study.

Table 5. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.503
Bartlett's Test of Sphericity	Approx. Chi-Square	19.527
	df	6
	Sig.	.003

Table 6. Communalities

	Initial	Extraction
Zscore (age)	1.000	.552
Zscore (BMI)	1.000	.631
Zscore (children)	1.000	.965
Zscore (smokercode)	1.000	.970

Note: Extraction Method: Principal Component Analysis.

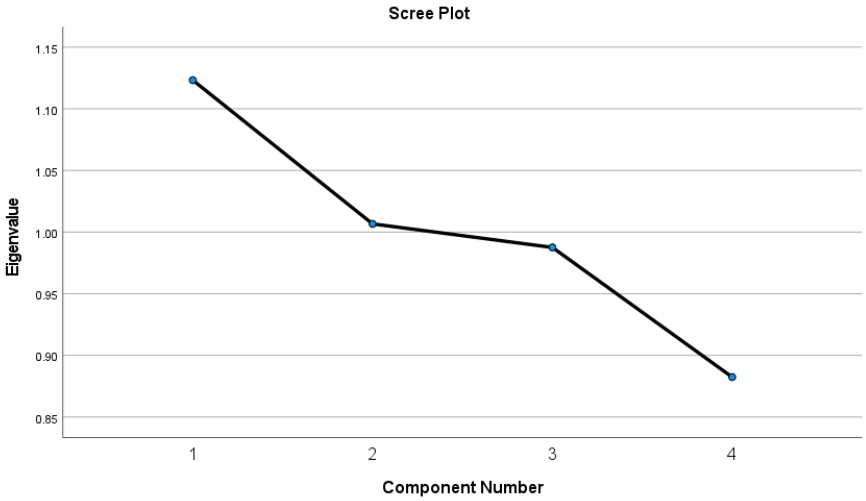

Fig. 2. Scree Plot

Table 7. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings	
	Total	% of Variance	Cumulative %	Total	% of Variance
1	1.123	28.082	28.082	1.123	28.082
2	1.007	25.167	53.249	1.007	25.167
3	.988	24.690	77.939	.988	24.690
4	.882	22.061	100.000		

Total Variance Explained				
Component	Extraction Sums of Squared Loadings	Rotation Sums of Squared Loadings		
	Cumulative %	Total	% of Variance	Cumulative %
1	28.082	1.109	27.730	27.730
2	53.249	1.004	25.111	52.841
3	77.939	1.004	25.098	77.939
4				
Extraction Method: Principal Component Analysis.				

Table 8. Component Matrix[a]

	Component		
	1	2	3
Zscore (age)	.737		
Zscore (bmi)	.683		-.405
Zscore (children)	.318	-.480	.796
Zscore (smokercode)		.877	.435

Extraction Method: Principal Component Analysis. [a]
a. 3 components extracted.

As shown in Table 5 rejecting the null hypothesis is that the correlation coefficient matrix is a unit matrix, that is, there is no correlation between variables, so the variables are considered to be related, that is, suitable for factor analysis.

For Table 6, we extracted the following three factors.

Factor 1- Whether or not the customer is a smoker

In Table 8 the items in factor 1 demonstrate significantly high loadings with variables such as age and the number of children. Two variables with positive loadings are extracted from one factor, and the positive loadings indicate the covariant proportionality of the two variables. As it pertains to the premium of life of smokers or not, the age and body mass index were .737 and .683, respectively. Given the nature of the variable in this factor, named the most vital primary factor is the physique factor.

Factor 2- The number of children the customer has

In Table 8 the variable of whether or not the customer is a smoker loadings of item in factor 2 is significantly high. Indeed, extracted positive loadings on factor 2 carry more weight as an aspect of the number of children the customer has, whether or not the customer is a smoker （.877） when predicting life insurance charges concerned. Based on the characteristics of the variables involved, this factor can be referred to as the family factor.

Factor 3- Body mass index (BMI)

Table 8 highlights that the variables of a customer's smoking status and the number of children have significantly high loadings in factor 3. The positive loadings suggest that these two variables are correlated with each other. When predicting life insurance charges, the Body Mass Index (BMI) of customers is influenced by whether or not the customer is a smoker (.435) and the number of children they have (.796) as important attributes. Considering the characteristics of the variables involved in this factor, it can be interpreted as a factor indicating the presence or absence of poor eating habits.

Comparative analysis

SPSS Comparative Analysis is a statistical technique that is used to compare two or more groups or variables with a focus on identifying differences or similarities between them. With SPSS software, comparative analysis can be done using a variety of Statistical analyses, including t-tests, ANOVA, and chi-square tests, which are employed to examine and evaluate the data. SPSS Comparative Analysis allows researchers to identify trends, patterns, and relationships between variables, which can help them to make informed decisions based on the data.

According to the characteristics of the database, the insurance charge in the database can be defined as Keys A (insurance charges≤10,000) and large insurance charge Keys B (insurance charge > 10,000). In this grouping, the large insurance cost and microinsurance cost were compared with the variables of age, BMI, gender, number of children the customer had, whether the customer smoked, and region. Since the three variables of age, BMI, and the number of children

the customer has are continuous variables, the total number of samples passing the normality test table is 1337, and the variables Age, BMI, and the number of Children have, all have a significance < 0.05 under the Tests of Normality (see Table 9), so they are not normally distributed. These three variables were analyzed using a Hypothesis Test (see Table 10). Since the three variables of sex, whether the customer smokes, and the customer's region are discrete, they are compared with large premiums and small premiums using the chi-square test (see Table 11). Since the Pearson chi-square > 0.05 under the Asymptotic Significance (2-sided) of the two variables SEX and Region in the chi-square test, it is not discussed that they do not significantly differ between large insurance charges and small insurance charges.

Table 9. Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	keys	Statistic	df	Sig.	Statistic	df	Sig.
age	A	.087	712	.000	.941	712	.000
	B	.160	626	.000	.892	626	.000
BMI	A	.035	712	.042	.992	712	.000
	B	.042	626	.010	.994	626	.016
children	A	.230	712	.000	.822	712	.000
	B	.267	626	.000	.819	626	.000

Note: a. Lilliefors Significance Correction

Table 10. Hypothesis Test Summary

	Null Hypothesis	Test	Sig. [a,b]
1	The distribution of age is the same across categories of keys.	Independent-Samples Mann-Whitney U Test	.000
2	The distribution of BMI is the same across categories of keys.	Independent-Samples Mann-Whitney U Test	.007
3	The distribution of children is the same across categories of keys.	Independent-Samples Mann-Whitney U Test	.882

Hypothesis Test Summary

	Decision
1	Reject the null hypothesis.
2	Reject the null hypothesis.
3	Retain the null hypothesis.

Note: a. The significance level is .050. b. Asymptotic significance is displayed.

Table 11. Descriptives

	keys		Statistic	Std. Error
age	A	Mean	33.09	.394
	B	Mean	46.16	.573
BMI	A	Mean	30.26035	.231730
	B	Mean	31.12181	.238681
children	A	Mean	1.08	.044
	B	Mean	1.12	.050

Table 12. Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	391.896[a]	1	.000		
Continuity Correction[b]	389.213	1	.000		
Likelihood Ratio	498.555	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	1338				

Note： a. 0 cells (.0%) have an expected count of less than 5. The minimum expected count is 128.19.

b. Computed only for a 2x2 table

As Table 10 shows Age=.000<.05 and BMI=.007<.05, both reject the null hypothesis that large insurance charges and microinsurance charges are evenly distributed, and the differences are significant. Further, according to the analysis of Table 11, the age of customers with an insurance charge of ≤ 10,000 is significantly greater than that of customers with an insurance charge of 10,000 >; BMI also has a greater impact on customers with ≤ insurance charge of 10,000 than customers with > insurance charge of 10,000. According to Table 10, the number of customers with children = .882>.05 shows that the null hypothesis of uniform distribution between large insurance charges and microinsurance charges is accepted, and the difference is not significant.

As shown in Table 12, the whether or not customers smoked variable is useful for Keys A (insurance charge≤ 10,000) versus Keys B (insurance cost >10,000) Pearson chi-square = 0 under Asymptotic Significance (2-sided). The results of .000<0.05 showed a significant difference between smoking and microinsurance charges for large insurance charges. People who smoke have a greater impact on large insurance costs.

4 Conclusion

The present analysis of the empirical literature on the affecting of factors for life insurance charges does not claim to be exhaustive but nevertheless affords an initial overview of the discussion status. In the analyzed study, a linear regression model was constructed to find the regression equation of variables for insurance charges and the correlation between variables. And through factor analysis, find out the degree of influence of each factor on insurance charges in the order of importance from largest to smallest. Finally, through comparative analysis, it is concluded that the older the customer, the more insurance charges will be paid, and the healthier

the customer, the less the expenditure on insurance charges; Customers who smoke pay more on insurance premiums.

The future disquisition needs to further deepen the study of this problem and put forward a new life insurance cost forecasting method. Divide data sets into groups with similar attributes through customer segmentation, for example, dividing customers into similar work history, medical history, and growth history to build a database, through which more accurate predictive models can be achieved to help life insurance customer datasets with different data mining methods. We are blessed with new opportunities and faced with new challenges that Data analytics is popular among companies worldwide. Big Data has contributed to the transformation of processes, organizations, and many aspects of the insurance sector. It is obvious that in order to be successful and competitive in the constantly evolving insurance market, insurance institutions must use Big Data. In the future, it should be expected that information collected with the use of new technologies will play an increasingly important role in the activities of insurers. In the future, accurate prediction of a person's one-year insurance charges and its influencing factors can be achieved by the internal optimization algorithm of artificial intelligence robots, called chatbots and how to combine artificial intelligence to further promote the development of the insurance industry is a problem worth studying.

References

[1]Sandeep Chaudhary. (2016) Consumer Perception Regarding Life Insurance Policies: A Factor Analytical Approach. Pacific Business Review International 9 (6), 52-61.

[2]Information Courtesy of USAA. (2022) 20 factors that affect property insurance rates. 20 Factors That Affect Property Insurance Rates | USAA

[3]Noorhannah Boodhun & Manoj Jayabalan. (2018) Risk prediction in the life insurance industry using supervised learning algorithms. Complex & Intelligent Systems 4, 145-154.

[4]Negi, M., & Kaur, Dr. P. (2010). A Study of Customer Satisfaction with Life Insurance in Chandigarh Tricity. Paradigm, 14(2), 29-44.

[5]Manuel (2013) Consumer Perception Regarding Life Insurance Policies: A Factor Analytical Approach. Pacific Business Review International 9 (6), 52-61.

[6]Singh S; Sirohi N &Chaudhary K. (2014) A Study of Customer Perception towards Service Quality of Life Insurance Companies in Delhi NCR Region. In: Global Journal of Management and Business Research. pp 19-32.

[7]Ganesh Dash. (2018) Determines of Life Insurance Demand: Evident from India Asia Pacific Journal of Advanced Business and Social Studies. ISBN (eBook) Publishing.

[8]Sindhuja R & Dr.M.P. Kumaran. (2021) A Study on Customer Buying Behaviour in Life Insurance Company with Special Reference to Coimbatore City. Peer Reviewed and Refereed Journal 4(5): 2380-2882.

[9]Bob Wakefield. Insurance. https://data.world/bob-wakefield/insurance.

[10] Szaniewski, Daniel. (2023) 11 Big data analytics in insurance. In: Lech,G., Jan, M. (Eds.), The Digital Revolution in Banking, Insurance and Capital Markets. Routledge Publishing Inc., New York.