The Application of Mathematical Statistics in Insurance

JianXin Zheng
School of Mathematical Sciences, TIANGONG University, Tianjin, China.
747158067@qq.com

Abstract
This article mainly uses principal component analysis, factor analysis, regression analysis and other statistical methods to analyze the effects of regional GDP, per capita GDP, per capita disposable income, resident RMB savings deposits, resident consumption levels, and the number of illiterate populations on the original insurance premium income. Although my country's premium income has grown rapidly in recent years, there is still a lot of room for development in the depth and density of insurance in my country, especially for people with different income levels and consumption levels, educated and uneducated people in different regions. There are big differences in demand, so at the end of this article, we will put forward some reasonable suggestions for the development of the insurance industry in various regions through the different degrees of influence of these factors.

Keywords
Principal component analysis, Regression analysis, Original insurance premium.

1. Introduction
The insurance industry has developed rapidly since its recovery in 1980 and has achieved remarkable results[1]. One of the fastest growing industries in China's national economy is the insurance industry. From the perspective of the overall environment, the economic situation at home and abroad is severe and complex. The real economy, financial market and consumer demand increase the difficulty of the insurance industry for stable operation of the insurance market through various channels. From the perspective of foreign conditions, China, as an open market, has huge potential and is very attractive to the international insurance market[2]. For the domestic market, although the international economic trend is severe and complex, the background of China's economic growth model is changing, and the economic growth rate has also become slow, China's insurance industry still maintains a stable development trend[3]. With the increase in people's income in recent years and the increase in the awareness of health, pension and other safety guarantees, insurance demand has become an indispensable part of people's daily consumption, and it occupies an increasingly important position in my country's consumption structure.

As one of the four pillars of the financial industry, the insurance industry is also an important part of the national economy[4]. Its growth and growth are of great significance to the healthy development of China's national economy.[5] Although China's premium income has grown rapidly over the years, the depth and density of insurance are still at the lowest level in the world. At the same time, the uneven regionalization of China's insurance market structure is very different. Therefore, studying the influencing factors of insurance premium income is helpful to the study of the development space of the insurance industry, and is also of great significance to the development of the insurance industry and macroeconomic development.

In recent years, my country’s premium income has grown rapidly, but people’s understanding of insurance differs greatly in its universality and importance.[6] In order to study what factors are related
to insurance premium income, I looked for some data and finally decided to analyze production in various regions. The six factors of total value, per capita GDP, per capita disposable income, resident RMB savings deposits, resident consumption level, and number of illiterate population have the impact on premium income. And used the principal component analysis, factor analysis and regression analysis learned to provide basis and data analysis for decision-making.

2. Introduction to Theoretical Part Section Headings

2.1 Related analysis
Correlation analysis is the basis of professional statistical analysis. It is a statistical method to explore the closeness of the relationship between various variables and is widely used. In statistical analysis, correlation coefficient is often used to quantitatively describe the linearity between two variables. Correlation analysis is used to explain the closeness between two variables, which reflects the degree of change of one variable when the value of another variable is controlled. The main purpose of correlation analysis is to study the degree of close relationship between variables, and to speculate based on sample data to determine whether the overall data is relevant. In statistical analysis, correlation coefficient is often used to quantitatively describe the linearity between two variables. For example, the relationship between family input and expenditure, and whether there is a relationship between the height of the child and the height of the parent.

The correlation coefficient is the square root of the coefficient in the regression equation. In the regression equation, the determined coefficient also reflects the closeness of the relationship between the two variables. However, the coefficient of determination is usually calculated after fitting the regression equation. If the relationship between the variables is not close, fitting the regression equation is meaningless. Therefore, correlation analysis is often performed before regression analysis. The non-discrimination of variables is a significant feature of correlation analysis, which shows that all variables are equally important.

2.2 Regression analysis
In quantitative analysis, you will always see a certain connection between variables. In order to understand the interaction between variables, we need to use correlation analysis and regression analysis. Regression analysis is an analysis method that is highly used in statistics and has a wide range of applications. It has been successfully applied in the fields of finance, economics, and medicine. It is used to study the statistical connection between things, the focus of the examination is on the quantitative changes between variables, and the regression equation is used to describe and reflect the relationship between variables.

Linear regression is a statistical method used to study the existence of a linear relationship between a dependent variable and one or more independent variables. In linear regression, if there is only one independent variable involved in regression, it is called linear regression analysis, and the result is called linear regression equation; if multiple variables are involved in regression analysis, then it is called multiple linear regression. In this article, we are using multiple linear regression.

Multiple linear regression model:
Suppose the relationship between the dependent variable and the independent variable, ..., is as follows:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \]

2.3 Principal component analysis
Principal component analysis was first proposed by Hotelling in 1933. Principal component analysis is based on the premise that a small amount of information is lost, using the idea of "diminishing dimensions" and taking several indicators as several comprehensive indicators as the main components. Each principal component is a linear combination of the original variables, and each
principal component is not related to each other, which makes the principal components have some superior performance than the original variables. Should continue to use other multivariate statistical methods to solve practical problems on the basis of principal component analysis, this is because principal component analysis cannot be used as the result of research.

3. Application

3.1 Regression analysis application

This article analyzes the influencing factors of the original premium income of various regions in 2013. Figure 3-1 Data source: China Statistical Yearbook.

Through filtering, six variables were selected.

Let the variable be the regional gross product (100 million yuan)-refers to the final results of the production activities of all the population in each region in a certain period of time. It can reflect the impact of premium income under large-scale production activities.

Per capita regional product (yuan)-refers to the regional GDP calculated by the average permanent population in a certain period of time. Reflects the impact of the average production level in a region on premium income.

Per capita disposable income (yuan)-is the income that residents can use for their free disposal. Because it is the most important determinant of consumer spending, it is often used to measure changes in a country's living standards.

Renminbi savings deposits for residents (100 million yuan)-is a part of the income that is not consumed by residents. It can well withdraw the relationship between income and expenditure of residents in a region.

Resident consumption level (yuan)—refers to the extent to which local residents meet the needs of people for survival, development and enjoyment in the process of consuming material products and labor services. It can be reflected in the quantity and quality of the material products and services consumed.

Illiterate population (number)-refers to the number of people without education in a region. It reflects the role of education in promoting insurance income.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Model</th>
<th>Non-standardized coefficient</th>
<th>Standard coefficient</th>
<th>t</th>
<th>Sig.</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>Standar d error</td>
<td></td>
<td></td>
<td>Zero order</td>
</tr>
<tr>
<td>1</td>
<td>(constant)</td>
<td>-1702413.616</td>
<td>625976.219</td>
<td>-2.720</td>
<td>.012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gross Regional Product (100 million yuan)</td>
<td>60.608</td>
<td>46.723</td>
<td>.244</td>
<td>1.297</td>
<td>.207</td>
</tr>
<tr>
<td></td>
<td>GDP per capita (yuan)</td>
<td>5.636</td>
<td>28.215</td>
<td>.030</td>
<td>.200</td>
<td>.843</td>
</tr>
<tr>
<td></td>
<td>Per capita disposable income (yuan)</td>
<td>275.524</td>
<td>116.236</td>
<td>.546</td>
<td>2.370</td>
<td>.026</td>
</tr>
<tr>
<td></td>
<td>Resident RMB savings deposits (100 million yuan)</td>
<td>210.274</td>
<td>65.310</td>
<td>.591</td>
<td>3.220</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>Resident consumption level (yuan)</td>
<td>-217.132</td>
<td>144.155</td>
<td>-.406</td>
<td>1.506</td>
<td>.145</td>
</tr>
<tr>
<td></td>
<td>Illiterate population (number)</td>
<td>544.602</td>
<td>257.127</td>
<td>.144</td>
<td>2.118</td>
<td>.045</td>
</tr>
</tbody>
</table>

Use SPSS software to perform regression analysis on the data and get Table 3-2. According to the constant term of the regression model given in Table 3-2, the partial regression coefficients of regional GDP, per capita GDP, per capita disposable income, resident RMB savings deposits, resident
consumption level, and number of illiterate population, are -1702413.616, They are 60.608, 5.636, 275.524, 210.274, -217.132, and 544.602. Among them, the constant term still represents the value of the dependent variable when the independent variable takes the value 0, but the constant term here obviously has no practical meaning. Therefore, according to the results shown in Table 1, the following regression model can be derived:

\[ y = -1702413.616 + 60.608x_1 + 5.636x_2 + 275.524x_3 + 210.274x_4 - 217.132x_5 + 544.602x_6. \]

3.2 Principal component analysis application

In an empirical study of a problem, in order to reflect the problem more comprehensively and more effectively, people often need to consider some related indicators, namely variables. This raises two questions: On the one hand, it is necessary to consider as many indicators as possible to avoid the omission of important information; on the other hand, the increase in the index increases the burden of research questions. In addition, because each indicator is a reflection of the same problem, it will cause overlap of information and cause collinearity between variables. Therefore, empirical researchers expect to use linear combinations to avoid the correlation between variables affecting the information collection of indicators.

Principal component analysis is the study of how to make several linear combinations of original variables to explain most of the information of the original variables. This was first proposed by Hotelling in 1933. When the first linear combination cannot extract more information, consider the second linear combination to continue the rapid extraction process, knowing that the extracted information is almost the same as the original metric. This is the basic idea of principal component analysis.

Principal component analysis was first proposed by Hotelling in 1933 to explore how to use a few linear combinations of original variables to explain most of the information contained in them. The basic idea of principal component analysis is that when the first linear combination cannot provide more effective information, the second linear combination is used for extraction until the extracted information is not much different from the original.

Generally, when principal component analysis can be used, fewer principal components can be used to obtain a small amount of information, and the corresponding principal components can be used as components to obtain random variables with smaller dimensions. Therefore, principal component analysis can retain most of the information of the original data while reducing the "dimensionality" of the data.

In principle, if there are n variables, then at most n principal components can be extracted, but if n principal components are extracted, the method is simplified and the actual meaning of the number of variables is simplified. Generally, it is sufficient to extract more than 90% of the information containing the first 2 to 3 main components.

The "information" of data is the limited amount of information provided by the variable or data when the variable only needs one piece of data. If the variable uses a series of different data, information such as the maximum value, minimum value and average value can also be obtained from it. The greater the difference of the variables, the more comprehensive the information they provide. Therefore, the information in principal component analysis is the variability of indicators, expressed in standard deviation or variance.

The mathematical model of principal component analysis:

\[
\begin{align*}
Y_1 &= u_{11}X_1 + u_{12}X_2 + \ldots + u_{1p}X_p, \\
Y_2 &= u_{21}X_1 + u_{22}X_2 + \ldots + u_{2p}X_p, \\
&\vdots \\
Y_p &= u_{p1}X_1 + u_{p2}X_2 + \ldots + u_{pp}X_p,
\end{align*}
\]
The output in Table 2 is the main component coefficient matrix, which can explain the loading of each principal component on each variable, and thus obtain the expression of each principal component.

### Table 2 Component score coefficient matrix

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gross Regional Product (100 million yuan)</td>
<td>.019</td>
<td>.386</td>
</tr>
<tr>
<td>GDP per capita (yuan)</td>
<td>.313</td>
<td>-.063</td>
</tr>
<tr>
<td>Per capita disposable income (yuan)</td>
<td>.311</td>
<td>-.042</td>
</tr>
<tr>
<td>Resident RMB savings deposits (100 million yuan)</td>
<td>.053</td>
<td>.358</td>
</tr>
<tr>
<td>Resident consumption level (yuan)</td>
<td>.311</td>
<td>-.030</td>
</tr>
<tr>
<td>Illiterate population (number)</td>
<td>-.202</td>
<td>.424</td>
</tr>
</tbody>
</table>

The principal component expression is:

\[
\begin{align*}
F_1 &= 0.019Y_1 + 0.313Y_2 + 0.311Y_3 + 0.053Y_4 + 0.311Y_5 - 0.202Y_6, \\
F_2 &= 0.386Y_1 - 0.063Y_2 - 0.042Y_3 + 0.358Y_4 - 0.030Y_5 + 0.424Y_6,
\end{align*}
\]

In the expression of the first principal component, the coefficients of, and are relatively large, which can be regarded as comprehensive indicators reflecting the per capita GDP, per capita disposable income, and residents' consumption level. In the second principal component expression, the coefficients of, and are relatively large, which can be regarded as a comprehensive indicator reflecting the regional GDP, resident RMB savings deposits and the number of illiterate population.

Principal component analysis is essentially a process of matrix transformation. For practical significance, we will perform factor analysis.

### 4. Conclusion and Suggest

From the final regression model, we can see that the consumption and quality of life of residents in various regions have affected the income of premiums. Although the six variables I chose have correlations, they are all positively correlated from the regression model, so we propose the following Suggest:

**Efforts to increase insurance awareness.** With the emergence of new forms of materials and commodities, people will also enjoy new risks. This requires people to establish and develop risk awareness in the new situation. The government can increase the scope and intensity of risk awareness education to make it a universal education, and it can also cultivate professional talents. Through training and publicity, the effect of popularization is an important issue.

**Diversified insurance products.** Since the reform and opening up, China's economic market has formed a diversified structure of goods. In high-consumption areas, the level of insurance consumption is not high. The diversified products will stimulate consumers to spend more on insurance according to different needs.

Finally, as the insurance industry continues to grow and develop, I believe that through continuous efforts, China's insurance industry must have a more brilliant future.

### References


