ISSN 2618-1584 Vol. 3, Issue 6: 21-26, DOI: 10.25236/FMSR.2021.030605

Analysis of Population Life Span Based on Grey Prediction

Yufeng Wu[1], Ruobing Zhang[2]

[1]*School of Mathematical Sciences, Chongqing Normal University, Chongqing 405400, China*
[2]*Faculty of Science, Tibet University, Lhasa 850000, China*

Abstract: *This paper studies the changing trend of population life expectancy and the impact of various influencing factors on life expectancy, and collects 8 influencing factor data from some provinces and cities in different locations across the country, and performs factor analysis on the 8 factors to determine their total Linear, classify the indicators with strong correlation into one category, make them a common factor, and name them. Then establish a multiple linear regression model of common factors and life expectancy; and use SPSS to fit the model to evaluate the accuracy of the built model. Finally, the gray prediction model is used to obtain the predicted value of 8 variables, and then the predicted value of the common factor is obtained, which is substituted into the regression equation to obtain the predicted life expectancy of the future population.*

Keywords: *life expectancy, grey prediction, factor analysis, multiple linear regression*

1. Introduction

According to research by archaeologists, the average life expectancy of Pekingese from 500,000 to 200,000 years ago was only about 17 years. After human beings entered a civilized society, the average life span has been extended. According to relevant data, the average life expectancy of ancient Greeks is estimated to be 20-30 years, ancient Romans are 15-30 years old, the average life expectancy of medieval England is estimated to be 33 years, and the average life expectancy of Americans in the early 20th century is estimated to be 48 years. About. According to statistics from the United Nations, by 1995, the average life expectancy in the world had reached 65 years. In 2007, Japan had the longest life expectancy in the world, with an average of 79 years for men and 86 years for women. Since 1978, China's society has developed rapidly and its economy has developed rapidly. The average life expectancy of the population has been continuously increasing from 30 or 40 years old in ancient times to 70 or 80 years old today, and the rate of increase of women is higher than that of men, which is in line with the general law of human life. Domestic scholar Zhao Tongqing [1] studied the effects of "death hormone", "senile hormone" and free radicals on human life span. Domestic scholar Cen Haiyan and others[2] studied the influence of physiological and psychological factors on life span. But they all only studied the impact of unilateral factors on life span, rather one-sided, and there is no research on the impact of multiple factors on life span.

This article studies the impact of multiple aspects on life span, such as medical care, education, environment, etc., and uses factor analysis on the collected data. That is to study the relationship between common factors and life span. Then establish the regression equation of the common factor and the population life expectancy, and use the gray prediction model to predict the future age.

2. Research object

In order to be as comprehensive as possible and truly reflect the national life expectancy, it is necessary to select representative indicators from various provincial and municipal data. For example: the number of medical and health institutions per 1,000 population (units) (x_1), the number of medical staff per 1,000 population (persons) (x_2), the average years of education (years) (x_3), the number of colleges and universities per 100,000 people (Person) (x_4), per capita consumption expenditure of urban residents (yuan/person) (x_5), per capita consumption expenditure of rural residents (yuan/person) (x_6), per capita living area in rural areas (m ²/person) (x_7), per capita water resources (m^2/person) (x_8) are used to judge life expectancy.

ISSN 2618-1584 Vol. 3, Issue 6: 21-26, DOI: 10.25236/FMSR.2021.030605

This article selects the corresponding index data of some provinces and cities from the statistical yearbook.

Python data processing: In recent years, more and more researchers have begun to use python to process data. It is simple and easy to learn, has good portability, scalability and embeddability, and has a wealth of libraries.

This article uses third-party data packages such as numpy and pandas to process the data.

Factor analysis: Factor analysis is the conversion of multiple measured variables into a few comprehensive indicators (or latent variables). It reflects the idea of dimensionality reduction, which brings together highly correlated variables through dimensionality reduction. Thereby reducing the number of variables that need to be analyzed, thereby reducing the complexity of problem analysis. The basic idea of the factor analysis method is to classify the observed variables, and classify them into the same category with higher correlation and closer connection, while the correlation between different variables is lower, then each type of variable actually represents a basic structure, the common factor.

Judgment of collinearity[3]*:* There are many indicators selected in this article, and the direct use of these variables to establish regression equations will reduce the prediction effect. Therefore, the use of factor analysis to achieve dimensionality reduction, extracting some relatively strong variables into common factors, and using common factors to establish prediction equations can make the prediction more accurate. The correlation coefficient matrix is used to verify whether there is correlation. If the correlation coefficient is greater than 0.3, it has strong correlation, that is, it has strong collinearity, which is suitable for factor analysis [4]. This article uses python to process the original data matrix and normalize it so that the processed data obeys a standard normal distribution with a mean of 0 and a standard deviation of 1. After the data is normalized, the matrix of the correlation coefficient of the original matrix is obtained.

Table 1: Correlation coefficient matrix

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	1	0.9256	0.7689	0.7459	0.6754	0.7121	0.2365	-0.1870
x_2	0.9327	1	0.7856	0.879	0.7845	0.7969	0.2609	-0.1587
x_3	0.7589	0.7690	1	0.769	0.8709	0.6509	0.3289	-0.6280
x_4	0.7850	0.8790	0.8750	1	0.7850	0.6890	0.6590	-0.2089
x_5	0.6590	0.7899	0.6825	0.7908	1	0.9087	0.6278	-0.1980
x_6	0.7090	0.7869	0.6780	0.7992	0.9076	1	0.7609	-0.2890
x_7	0.2189	0.2689	0.3179	0.3609	0.6245	0.7267	1	-0.1570
x_8	-0.1546	-0.1679	-0.6124	-0.203	-0.198	-0.214	-0.1450	1

It is known from the Table 1 that most of the absolute values of the correlation coefficients among the eight variables are greater than 0.3, which has a strong correlation. That is suitable for factor analysis for dimensionality reduction.

3. Extract common factors

Calculate the eigenvalues and eigenvectors of the correlation coefficient matrix, use the variable names and eigenvalues to reconstitute a table, use this table to confirm the number of common factors, and draw a graph of the eigenvalue changes of the variables.

Table 2: Characteristic value data

variable name	Eigenvalues
Number of medical and health institutions per thousand population (units)	5.688235
Number of medical staff per thousand population (person)	1.496217
Average years of education (years)	1.251124
Number of students in colleges and universities per 100,000 people (persons)	0.646407
Per capita consumption expenditure of urban residents (yuan/person)	0.361241
Per capita consumption expenditure of rural residents (yuan/person)	0.268689
Per capita living area in rural areas (m^2/person)	0.067464
Water resources per capita (m^2/person)	0.023358

ISSN 2618-1584 Vol. 3, Issue 6: 21-26, DOI: 10.25236/FMSR.2021.030605

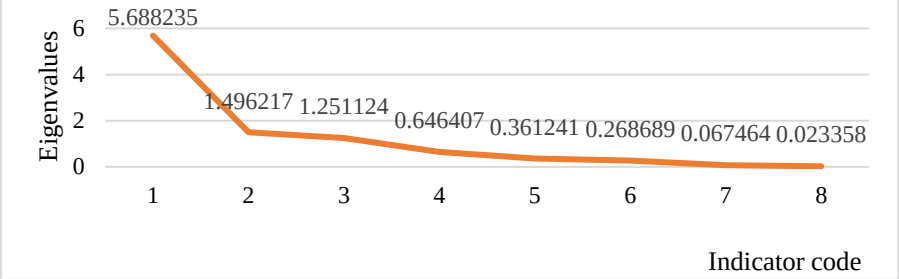

Figure 1: Variation of eigenvalues

It can be seen from Figure 1 and Table 2 that the eigenvalue difference starts to decrease after x_3, the initial eigenvalue is accumulated and the eigenvalue sum is quotient, and the eigenvalue is accumulated after 3 cycles, and the eigenvalue accumulated value is 84.365 at this time. It shows that the first three factors can intensively explain the selected eight indicators, and they contain most of the information in the eight variables, so the original variable group can be collectively expressed as three common factors after factor analysis.

4. Factor naming

From the factor loading matrix, it can be found that the load coefficient on the common factor has a large gap. In order to redistribute the proportion of variance explained by each factor so that the load coefficient is closer to 0 or 1, then the variables can be better explained and named. Therefore, it is necessary to perform factor rotation on the factor loading matrix by changing the position of the coordinate axis. The rotated factors do not change the degree of fit of the model to the data, nor does it change the variance of the common factors of each variable, making the factor structure simpler, so that the common factors can be better extracted for naming. This paper adopts the maximum variance orthogonal rotation method, defines a function varimax to achieve this function, and gets the rotated component matrix.

Draw the following conclusions from the rotated component matrix:

1) The number of medical and health institutions per thousand population, the number of medical staff per thousand population, the average years of education, the number of students in colleges and universities per 100,000 people, the loading ratio of these four factors on factor 1 is on the other two common factors both are bigger. These four factors mainly explain the level of health care and education, and can be named as livelihood factors.

2) The three factors of per capita consumption expenditure of urban residents, per capita consumption expenditure of rural residents, and per capita living area in rural areas have a larger load on factor 2 than on the other two common factors, which mainly explain people's daily life. They can be named as life factor.

3) The per capita water resources have a greater load on factor 3, which mainly explains the living environment conditions, which can be named environmental factors.

In order to obtain the common factor equation, it is necessary to calculate the score of each factor on the common factor to obtain the factor score matrix:

Table 3: Factor scoring matrix

	Common factor 1	Common factor 2	Common factor 3
Number of medical and health institutions per thousand population (units)	.325	-.152	-.119
Number of medical staff per thousand population (person)	.295	-.101	-.088
Average years of education (years)	.245	-.147	.243
Number of students in colleges and universities per 100,000 people (persons)	.248	-.068	-.014
Per capita consumption expenditure of urban residents (yuan/person)	.034	.241	-.048
Per capita consumption expenditure of rural residents (yuan/person)	.015	.275	-.053
Per capita living area in rural areas (m^2/person)	-.235	.506	-.055
Water resources per capita (m^2/person)	-.040	.121	-.539

From Table 3, the livelihood factors, living factors, and environmental factors obtained through factor analysis are labeled as F_1, F_2, and F_3, respectively. Through factor scores and standardized index data,

ISSN 2618-1584 Vol. 3, Issue 6: 21-26, DOI: 10.25236/FMSR.2021.030605

a common factor equation is established.

F_1=0.325x_1+0.295x_2+0.245x_3+0.248x_4+0.034x_5+0.015x_6-0.235x_7-0.040x_8

F_2=0.152x_1-0.101x_2-0.147x_3-0.0688x_4+0.241x_5+0.275x_6+0.506x_7+0.121x_8

F_3=-0.119x_1-0.088x_2+0.243x_3-0.014x_4-0.048x_5-0.053x_6-0.055x_7-0.539x_8

4.1. Multiple linear regression

Establish a multiple linear regression model of common factor and life expectancy [5]. There is no significant correlation between F_1, F_2, and F_3. Multiply the standardized index data obtained above and the factor score matrix to obtain the data matrix of the three common factors of the data used this time.

Table 4: Multiple linear regression data table

Province	Life expectancy	F_1	F_2	F_3
Beijing	80.18	3.511364	0.653314	-0.02002
Heilongjiang	75.98	0.469237	-0.87551	0.414121
Inner Mongolia	74.44	0.195224	-0.54301	0.590764
Hebei	74.97	-0.46426	-0.37409	1.014856
Shanghai	80.26	2.191621	2.243697	-0.12216
Zhejiang	77.73	-0.56733	2.930858	2189
Anhui	75.08	-0.81864	0.091782	0.249651
Fujian	75.76	-0.58371	0.916019	0.08658
Chongqing	75.70	-0.69812	0.436968	0.766764
Sichuan	74.75	-0.52751	-0.08644	-0.05642
Yunnan	69.54	-0.58202	-0.51873	-0.32351
Tibet	68.17	-1.13008	-0.02409	-4.46028
Shaanxi	74.68	0.320266	-0.57036	0.248084
Gansu	72.23	-0.15911	-0.97314	-0.31946
Ningxia	73.38	-0.3667	-0.50068	1.328649
Xinjiang	72.35	0.661698	-1.15638	-0.46084

Use life expectancy as the dependent variable, F_1, F_2, F_3 as independent variables. Use spss software to fit the regression model.

Table 5: Model summary

Model	R	R^2	Adjusted R^2	Standard skewness error
1	0.882	0.779	0.754	1.36262

R^2 is the ratio of the regression sum of squares to the total deviation of squares. The larger the ratio, the better. After adjustment, R^2 takes into account the number of regression independent variables, and there will be no large differences due to changes in the number of independent variables. After adjustment, R^2 is 0.754, indicating that the independent variables can explain 75.4% of the dependent variables. The difference indicates that the model is more appropriate.

Table 6: Coefficient

Model	Non-standardized coefficient		Standardization factor	T	Significance
	B	Standard error	Beta		
C	74.906	.245		306.073	.000
F_1	1.681	.249	.612	6.760	.000
F_2	1.290	.249	.469	5.182	.000
F_3	1.179	.249	.429	4.741	.000

Establish a multiple linear regression model through non-standardized coefficients:

$$Y=74.906+1.681F_1+1.290F_2+1.179F_3$$

4.2. Grey prediction[7, 8]

This article only collects data from some provinces and cities, with few data samples and no regularity. This article needs to predict the long-term life expectancy of the population. Therefore, this paper adopts

ISSN 2618-1584 Vol. 3, Issue 6: 21-26, DOI: 10.25236/FMSR.2021.030605

the gray prediction model to process small and irregular data, and to quickly predict the long-term population life.

Currently, the most widely used gray prediction model is the GM(1,1) model. The GM(1,1) model is based on a random original data sequence. A new data sequence formed after accumulation over time can be presented as a rule. The solution of the first-order linear differential equation is approximated.

For the original data column

$$X^{(0)} = \left[x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)\right]$$

Do an accumulation of 1-AGO once, and get the sequence:

$$X^{(1)} = \left[x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n)\right]$$

In,

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), k = 1, 2, \dots, n$$

Generate the adjacent mean value for the sequence $X^{(1)}$, and get the sequence:

$$Z^{(1)} = \left[z^{(1)}(2), z^{(1)}(3), \dots, z^{(1)}(n)\right]$$

In, $z^{(1)}(k)=0.5(x^{(1)}(k) + x^{(1)}(k-1)), \mathrm{k} = 2, 3, \dots, \mathrm{n}$

Let $x^{(0)}(k)$ and $z^{(1)}(k)$ as shown above, call $x^{(0)}(k) + az^{(1)}(k) = b$ is the gray differential equation of the GM(1,1) model. The parameter a is the development coefficient, and b is the gray effect amount.

$$\hat{a} = (a, b)^T = (B^T B)^{-1} * B^T * Y$$

In, $Y=\begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \dots \\ x^{(0)}(n) \end{bmatrix}$, $B=\begin{bmatrix} -z^{(1)}(2) \\ -z^{(1)}(3) \\ \dots \\ -z^{(1)}(n) \end{bmatrix}$

Corresponding to the gray differential equation of the GM(1,1) model, there is the whitening differential equation of the GM(1,1) model:

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = b$$

Solving the whitening differential equation, the GM(1,1) prediction model can be obtained as:

$$\hat{x}^{(1)}(k) = \frac{b}{a} + \left(x^{(0)}(1) - \frac{b}{a}\right) e^{-a(k-1)}, k = 2, 3, \dots, n$$

Through the first-order cumulative reduction of $\hat{x}^{(1)}(k)$, the final predicted value formula is obtained, Figure 2 shows the expected data of various indicators in recent and future years.

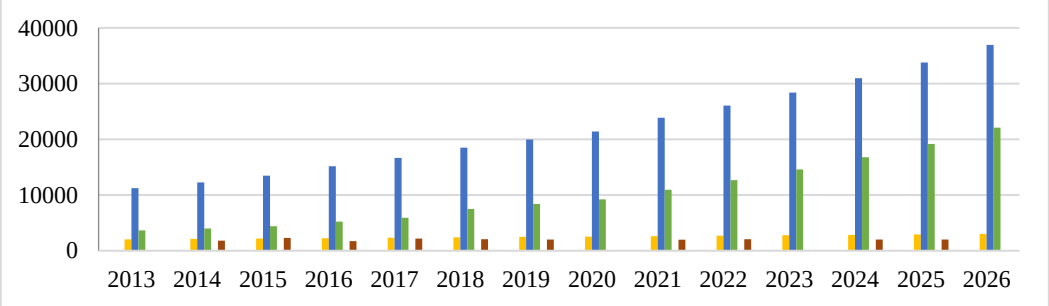

Figure 2: Index data

The Table 7 shows the simulated values of population life expectancy

ISSN 2618-1584 Vol. 3, Issue 6: 21-26, DOI: 10.25236/FMSR.2021.030605

Table 7: Life expectancy simulation values

Years	Actual value	Analog value	Relative error
2015	72.95	72.84531	0.14%
2016	74.83	73.09799	2.31%
2017	76.34	75.905	0.49%

It can be seen from Table 7 that the simulated value error calculated by the gray prediction regression model is relatively small and has a certain reference. The predicted life expectancy for the next few years is subsequently calculated.

Table 8: Population life expectancy

Years	2022	2023	2024	2025	2026
Life expectancy	77.81	78.30	78.91	79.64	80.38

In Table 8, the life expectancy of the population in the next five years is given.

5. Conclusion

This paper collects 8 representative indicators and collects their specific values. However, in order to make the prediction results more reasonable, the 8 indicators are judged for collinearity, and the indicators with strong correlation are classified into one category. In this way, on the one hand, the number of variables is reduced, making the model more concise and clear, and on the other hand, the amount of calculation is reduced, making the prediction results more accurate. After categorizing the indicators, a multiple linear regression model is established for the new variables and life, and the linear regression equation between the life and the new variables is obtained, and the SPSS software is used to fit the equation. If the error is small, the equation is in line with the actual situation. As for the life expectancy, this paper adopts the gray prediction model, which uses fewer and irregular samples for prediction, which can predict the result in a shorter time and save time and cost.

References

[1] Zhao T Q. Analysis of human life span [J]. Journal of Chongqing Three Gorges University, 2001, 4(17): 1008-4347

[2] Cen H Y, Zhang Y Q. Research progress on factors related to longevity. Guangxi Medicine, 2018, 40 (12).

[3] Fan S G, Xi S J. Comparison of multiple collinear methods in the multiple linear regression model—Taking the impact of population migration on education resource model as an example [J]. Science and Technology Wind, 2019, (23).

[4] Shao H M, Shao X Y. Optimization analysis of the company's financial competitiveness evaluation index system based on factor analysis—taking real estate listed companies as an example [J]. China Collective Economy, 2021, (27).

[5] Chen X J, Ji F X. Comprehensive evaluation of customer relationship management, internal control and corporate M&A performance measurement-based on multiple linear regression model analysis [J]. Management Review, 2021, 33(08).