Analysis of the Influence Degrees of Various Factors in the Verification of China's Pensions Based on Statistical Technology

Jiamei Sun*, Xi Liu

State Grid Jilin Electric Power Co. Ltd., Changchun, 132200, China

*Corresponding author's e-mail: 68633404@qq.com

Abstract. According to the basic pension approval policy, there are numerous factors affecting the level of pension, and it is impossible to judge specifically which factor has a greater or lesser impact through subjective judgment. This paper adopts the factor analysis method to analyze the actual pensions of retired workers and related data by using the important indicator screening model and linear regression model in statistics, to analyze the importance of the influencing factors of the basic pension approval, and to quantitatively compare the degree of influence of each factor on the level of basic pension.

Keywords: influencing factors, importance analysis, regression analysis

1 Introduction

According to the approved policy of basic pension, there are many factors affecting the level of pension, and it is impossible to judge specifically which factor has a large or small impact through subjective judgment. Three hundred and seventy employees who retired between 2019 and 2023 in four regions, namely Beijing, Jilin, Henan and Shaanxi, were selected as samples, and the basic pensions, personal information such as time of joining the workforce, time of retirement, and time of enrollment in the insurance program, as well as data on the factors influencing the approval of the basic pension, such as the contribution base, the crediting base, the number of years of contribution, the index of the contributory wage, and the transition coefficient, were collected from the abovementioned employees.[1]Using the basic pension as the dependent variable and the influencing factors as the independent variables, the importance of the approved influencing factors of the basic pension and the degree of influence were analyzed by using the importance indicator screening model and the linear regression model in statistics.[2]

To ensure the reasonableness and accuracy of the statistical analysis, nine influencing factors that are differentiated and quantifiable were selected as independent variables for this analysis, including the calculation base, the average value of the contribution base in previous years, the average value of the average social wage in previous years, the index of the deemed contribution wage, the actual number of years

H. Sun et al. (eds.), *Proceedings of the 2025 10th International Conference on Social Sciences and Economic Development (ICSSED 2025)*, Advances in Economics, Business and Management Research 332,
https://doi.org/10.2991/978-94-6463-734-2_30

of contribution, the deemed number of years of contribution, the transition coefficient, the number of months of calculation, and the amount of storage in the individual account.

2 Analysis of the Importance Ranking of Influencing Factors

Variable importance ranking is crucial in data analysis to help identify which features have a significant impact on the model's predicted results, thus improving the model's interpretability and accuracy. By ranking, the model can be streamlined to remove unimportant features and reduce the risk of overfitting, while improving the efficiency of the model's operation.

In this analysis, LightGBM is chosen as the model for screening the importance indicators of the influencing factors, and the independent variables that have the greatest influence on the dependent variable are accurately identified and ranked by machine learning algorithms. The reason for selecting LightGBM as the model choice for importance metrics screening is mainly due to its efficiency and excellent performance.The principle of LightGBM screening importance metrics mainly relies on the feature importance assessment mechanism in its decision tree algorithm. Specifically, it calculates the information gain to determine the contribution of each feature to the target variable when the node splits, and the larger the gain, the more important the feature is. Meanwhile, the frequency of features being used to split nodes and the coverage of participating samples also reflect the importance of features. In addition, LightGBM can be combined with SHAP values for more detailed feature interpretation to quantify the impact of each feature on the final prediction. These methods enable LightGBM to effectively identify key features, providing a solid foundation for subsequent data analysis and model optimization.

Using Python programming language, after importing the above independent and dependent variable data into the LightGBM model, the importance indicator ranking is obtained.

Among the nine factors affecting the approval of basic pensions, the importance scores of four factors, namely, the number of years of deemed contributions, the amount of personal account storage, the actual number of years of contributions, and the average value of the contribution base over the years, are significantly higher than the other factors, and they have a higher degree of influence on the level of pensions, while the remaining factors, though they also have an influence on the level of pensions, have a relatively lower degree of influence(see Table 1 below).

Table 1. LightGMB Importance Indicator Ranking and Analysis Table of Importance Indicators

Serial No.	Independent variable	Importance score	On the level of pensions Degree of impact
1	the deemed number of years of contribution	309	Maximum
2	the amount of storage in the indi-	236	Greater

	vidual account		
3	the actual number of years of contribution	171	Major
4	the average value of the contribution base in previous years	133	Great
5	the number of months of calculation	74	General
6	calculation base	49	Relatively low
7	the average value of the average social wage in previous years	46	Low
8	the index of the deemed contribution wage	28	Extremely low
9	the transition coefficient	27	Minimal

3 Linear Regression Analysis of Influencing Factors

After assessing the importance of the influencing factors, it was determined that all of the above independent variables have a certain degree of importance, so all of the above independent variables were included in the fitting of the linear regression model. However, there are significant correlations between some of the independent variables, or the values are too single, which may cause the regression results of some variables in the model to be insignificant or contrary to the theoretical expectations: for example, the contribution base is generally determined based on the social wage, so the average value of the social wage of the past years is highly correlated with the average value of the contribution base of the past years; the indexes of the deemed contribution wages in Beijing, Henan, and Shaanxi are 1, which is a too single value; and the storage amount of the individual account is accumulated from the contributions of the past years, so it is highly correlated with the average value of the contribution base of the past years. The amount of personal account storage is accumulated by the contributions in the past years, so it is highly correlated with the average value of the contribution base in the past years. Therefore, in order to improve the simplicity and explanatory power of the model, the regression analysis excludes three independent variables: the average value of the average social wage in the past years, the index of the deemed contribution wage and the amount of personal account reserve.[3]

Ordinary Least Squares (OLS) was chosen for this regression analysis and the results of the regression model obtained are presented in the table 2 below:

Table 2. Linear regression results

Meaning of independent variables	Regression coefficient analysis
intercept term	-8916.9528*** (0.000)
calculation base (yuan/month)	0.3742*** (0.000)
average value of the contribution base in previous	0.5226***

Meaning of independent variables	Regression coefficient analysis
years (yuan/month)	(0.000)
actual number of years of contribution (year)	276.0824*** (0.000)
deemed number of years of contribution (year)	289.2015*** (0.000)
transition coefficient (%)	$1.956e^{+05}$*** (0.000)
number of months of calculation (month)	-11.9795*** (0.000)
R^2	0.926
adjusted R^2	0.925
F-value	757.5

3.1 Overall Analysis of the Model

Ordinary Least Squares (OLS) is a method of estimating regression coefficients by minimizing the sum of squares of the residuals, which is effective in estimating the parameters of a linear regression model and the results are easy to interpret, and also provides Best Linear Unbiased Estimation (BLUE), which means that it provides the smallest possible variance under certain assumptions that are satisfied, thus improving the accuracy and reliability of the predictions.

In the table *** indicates that the coefficient estimate is significant at the 1% level, ** indicates that the coefficient estimate is significant at the 5% level, and * indicates that the coefficient estimate is significant at the 10% level; the smaller the level of significance, the lower the likelihood that the test conclusions will make an error.

The value within () is the p-test value, which is less than the significance level, indicating that the variable is significant.

3.1.1 Goodness of Fit of the Regression Model

R^2 and adjustedR^2: R^2 is a statistic that measures the degree of fit of a regression equation and reflects the proportion of the variance in the dependent variable that is explained by the estimated regression equation. Adjusted R^2 is similar to R^2, but adjusted R^2 takes into account the effects of both sample size and the number of independent variables in the model. The adjusted R^2 is usually used in multiple regression for the evaluation of the degree of fit.

Translated with DeepL.com (free version)R^2 is 0.926, and adjustedR^2 is 0.925, indicating that the model fits well, and up to 92.5% of all changes in the dependent variable can be explained by changes in the independent variable. [4]

3.1.2 Linear Relationship Test

F-value: The F-value tests the significance of the linear relationship between the dependent variable and all independent variables, also known as the overall significance test.

The F-value is 757.5, indicating that the linear relationship between the dependent variable and all the independent variables is significant.[5]

3.2 Analysis of Regression Coefficients

The regression coefficient of an independent variable indicates the average amount of change in the dependent variable for every 1 unit change in that independent variable, holding other variables constant.[6]

3.2.1 Calculating Base: the coefficient is 0.3742, i.e., for every 1,000 yuan/month increase in the calculating base, the average basic pension increases by 370.42 yuan/month.

3.2.2 The Average Value of the Contribution Base in Previous Years: the coefficient is 0.5226, i.e., for every increase of 1,000 yuan/month in the average value of the contribution base in previous years, the basic pension will increase by 522.60 yuan/month on average.

3.2.3 Actual Contribution Period: the coefficient is 276.0824, and for every increase of 1 year in the actual contribution period, the basic pension will be increased by an average of 276.08 yuan/month.

3.2.4 Deemed Contribution Period: the coefficient is 289.2015. For every additional year of deemed contribution period, the basic pension will increase by an average of 289.20 Yuan/month.

3.2.5 Transitional Coefficient: the coefficient is $1.956e^{+05}$, and for every 0.1% increase in the transitional coefficient, the basic pension increases by an average of 195.60 yuan/month.

3.2.6 Number of Months of Accrual: the coefficient is -11.9795, and for every increase of 1 month in the number of months of accrual, the basic pension is reduced by an average of 11.98 yuan/month.

4 Conclusion

The four factors mentioned above that have a high degree of influence on the level of pensions are all related to the retirees' own conditions, which further confirms that the degree of influence of their own conditions, such as the number of years of contributions, the age of retirement, and the contribution base, on the level of pensions generally exceeds that of the method of calculation and regional factors.[7]

Reference

1. Wenling Liu, Fengmin Xu, Kui jing and Ziyue Hua, Should the Occupational Pension Plans' Investment be Long-Term or Short-Term? Evidence from China, Computational Economics(2024), DOI: 10.1007/s10614-024-10677-3
2. Nelly Claire Muntalima, Adam Silumbwe, Joseph Mumba Zulu, Chris Mweemba and Peter Hangoma, Collaborative governance of an integrated system for collecting contributions for social health insurance, pension, and taxes from the informal sector: a synthesis of stakeholder perspectives, Muntalima et al. BMC Health Services Research (2024) 24:1253 DOI: 10.1186/s12913-024-11634-4
3. Meng Wang, Ming-liang Shu, Jian-jun Zhou, Si-xin Wu & Min Chen, Least Square Estimation for Multiple Functional Linear Model with Autoregressive Errors, Acta Mathematicae Applicatae Sinica, English Series, Volume 41, pages 84–98, (2025), DOI: 10.1007/s10255-024-1143-2
4. Scott H. Koeneman & Joseph E. Cavanaugh, A novel bootstrap goodness-of-fit test for normal linear regression models, AStA Advances in Statistical Analysis, November 2024, DOI: 10.1007/s10182-024-00517-y
5. Jianghao Li, Shizhe Hong, Zhenzhen Niu and Zhidong Bai, Test for high-dimensional linear hypothesis of mean vectors via random integration, Statistical Papers, Volume 66, article number 8, (2025) DOI: 10.1007/s00362-024-01624-3
6. Jin Hee Yoon, Dae Jong Kim and Yoo Young Koo, Novel Fuzzy Correlation Coefficient and Variable Selection Method for Fuzzy Regression Analysis Based on Distance Approach, International Journal of Fuzzy Systems, DOI: 10.1007/s40815-023-01546-6
7. Olga M. Fuentes, Richard K. Fullmer and Manuel García-Huitrón, A sustainable, variable lifetime retirement income solution for the Chilean pension system, The Geneva Papers on Risk and Insurance - Issues and Practice (2024) 49:234–258, DOI: 10.1057/s41288-024-00315-2