Analysis on HR management strategy based on principal component and multivariate regression model

Yayue Fang
Shandong Management University, Jinan City, China

Abstract. Regression analysis is a predictive modeling technology in big data analysis. In the field of statistics, regression analysis is an analytical method to determine the quantitative relationship between two or more variables. As one of the regression analysis methods, the multiple regression model can achieve the purpose of outcome prediction and discriminant analysis in the fields of data mining and economic forecasting. Especially in the case analysis with many influencing factors and complex conditions, the model is established from a linear perspective, and the influence of various factors on the output variables is analyzed through the weights in the obtained results. Combining multivariate analysis with principal component analysis, the main influencing factors can be screened out before correlation analysis, and the linear correlation analysis environment can be filtered to make the linear results more accurate and unsusceptible to extreme samples. This paper collects the data of national enterprise income and its related influencing factors from 2010 to 2020, and establishes a multiple regression model based on principal component analysis to study human resource management in China.

1. Introduction
With the structural reform of China's economy, though domestic enterprises have developed rapidly, it's also ushered in new opportunities and challenges. According to statistics, China's institutions have developed to more than 1.4 million, employees have exceeded 32 million, the introduction and management of talent will be closely related to the survival and development of enterprises. In the current era of knowledge informatization, human resources are particularly important. Many scholars have analyzed the current situation of human resources management in China, and in view of the existing problems in enterprise management, quantitatively analyzed human resources by using different algorithms, so as to formulate optimization strategies and ensure that China’s reform and development are always on the right path [1-2]. In 2006, Xie Lin and Du Gang used Delphi method and AHP method to establish a multi-level and multi-index evaluation model to discuss the flexible management of human resources. The model was applied to the evaluation of the flexibility of human resources management in a large hospital in Zhejiang Province. The comprehensive flexibility evaluation values of A, B, C, D and E were 4.44,4.20,4.25,4.92 and 3.74, respectively, indicating that the functional flexibility strategy of hospitals is particularly important compared with high-tech enterprises and iron and steel plants [3]. In 2010, Zhou Liqun and Xia Liangke used analytic hierarchy process and standard difference method to measure the degree of regional economic integration of Beijing-Tianjin-Hebei, Yangtze River Delta and Pearl River Delta, and obtained the values of...
economic integration indicators of the three major economic circles: from the point of view of the arithmetic average from 1989 to 2007, the Beijing-Tianjin-Hebei region was the lowest in the three economic circles, with an average of 0.461, and the Pearl River Delta was the highest, with 0.525, followed by the Yangtze River Delta and 0.478, indicating that the Pearl River Delta region had the lowest degree of economic integration. In 2014, Gao and Hua carried out empirical research on the spatial economic growth model with the help of the Lucas mechanism and the Nielsen-Phelps mechanism. The results of the Lucas mechanism cannot prove that human capital at all levels promotes economic growth, but the promotion of economic growth under the Nielsen-Phelps mechanism is confirmed. The AdjR2 of models 1 to 4 is 0.7949, 0.794, 0.7930 and 0.7929, respectively [4]. In 2021, Chen Jiahui used the coupling coordination model to study the interactive performance of human capital and economic development in the Yangtze River Delta, and divided the coupling classification of human resources system and economic system from (0, 1), as well as the coordination classification of human resources system and economic system, analyzed the coupling degree and coordination degree of 29 cities in the Yangtze River Delta region, and concluded that the coupling degree of most cities is located in (0.3, 0.5), that is, it is still in the initial stage of the grinding stage and the high coordination stage, and the coordination degree is located in (0.4, 0.5), that is, the coordination level develops from serious or moderate imbalance to intermediate and good coordinated development. It is concluded that there is a certain gap in regional economic development and there is still much room for development. However, in the existing literature at home and abroad, the multiple regression model is less used in the quantitative analysis of human resources, and the principal component analysis is used to screen the factors that affect human resource management. The principal component analysis can give the factors that have a relatively large impact on the target from a quantitative point of view, and give a more intuitive result. Therefore, this paper will analyze human resource management strategies based on the principal component analysis and the multiple regression model.

2. Algorithm and principle

Multiple regression analysis, as a generalized linear regression analysis model, is commonly used in data mining, economic forecasting and other fields. Through the statistical analysis of the linear or nonlinear relationship between the data of multiple influencing factors in the case, similar to Logistic regression, the model form has the structure of formula 1.

\[ y = ax + b \] (1)

Where a, b are the parameters to be calculated. Multiple linear regression is to build the basic model from the perspective of linear system, combined with the analysis results can establish a linear relationship, as formula 2.

\[ Z = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n + \cdots \] (2)

In the expression, \(x_1 \sim x_n\) is the influence factor of the output variable. \(a_0 \sim a_n\) is the regression coefficient of each influencing factor. The larger the absolute value is, the more obvious the influence of output variables is, and the larger the proportion of independent variables is.

In multivariate analysis, the determination method of regression coefficient is usually maximum likelihood method. The maximum likelihood method determines a certain probability distribution of random samples by observing the results through several experiments. According to the overall distribution, the likelihood function is \(L(x_1, x_2, ..., x_n; \Theta_1, \Theta_2, ..., \Theta_k)\), the solving steps are:

a. When the overall X is discrete, the probability distribution is listed as follows.

\[ P(X = x) = p(x, \Theta) \] (3)

Therefore,

\[ L(\Theta) = L(x_1, x_2, ..., x_n; \Theta) = \prod_{i=1}^{n} p(x_i, \Theta) \] (4)

b. When the overall X is continuous, its probability density function is \(f(x, \Theta)\), where \(\Theta\) is an unknown parameter.

Therefore,

\[ L(\Theta) = L(x_1, x_2, ..., x_n; \Theta) = \prod_{i=1}^{n} f(x_i, \Theta) \] (5)
The logarithm of likelihood function $L(\theta)$ is taken and arranged.
c. When $\theta$ is differentiable, find its derivative
\[ \frac{\partial \ln L}{\partial \theta_i} = 0 \] (6)
d. Solving the likelihood equation
Figure 1 shows the flow chart of the maximum likelihood method.

![Flow chart of maximum likelihood method](image)

**Figure 1.** Flow chart of maximum likelihood method.

3. Experimental results and discussion

3.1. Design experiment
In view of the human resources management and enterprise performance is closely related, this paper selects the enterprise annual income as the evaluation of human resources management strategy is scientific and effective basis, as the output variable. At the same time, according to the survey of the current situation of human resources in enterprises, the common problems are mainly divided into the following three aspects: backward personnel management concept, inhuman treatment of employees and unscientific post management. In reflecting the unit personnel management concept is backward or not, select the annual training times of enterprise management, which can reflect whether the enterprise learn advanced management ideas in time ; The average age of management can reflect whether enterprises attach importance to young talents and accept new ideas, but rely on qualifications and experience ; In terms of reflecting the humanization of employee treatment, the average weekly overtime, employee turnover rate and annual average overtime cost are selected to quantify. In terms of whether the job management is scientific, the annual turnover rate of employees, the professional counterparts rate of employees, the average recruitment salary of current students, and the average annual salary of senior talents (master’s degree and above) are selected to reflect.

Through the above analysis, this paper statistic the data of the above influencing factors and output variables from 2014 to 2020. The variable numbers are shown in Table 1, and the statistical data are shown in Table 2.
Table 1. Variable naming table.

| variable | meaning                                           |
|----------|---------------------------------------------------|
| Y        | Annual average total income of enterprises (yuan) |
| $X_1$    | Annual training times of enterprise management (times/year) |
| $X_2$    | Average age of management (years)                  |
| $X_3$    | The average weekly overtime of employees (week/hour) |
| $X_4$    | Annual staff turnover rate (%)                     |
| $X_5$    | Average weekly overtime (weekly/time)              |
| $X_6$    | Staff Professional counterparts (%)                |
| $X_7$    | Average Recruitment Pay for Freshmen (Yuan/Month)  |
| $X_8$    | Average annual salary of senior talents (Yuan/month) |

Table 2. Data on Enterprise Annual Income and Its Influencing Factors from 2010 to 2020.

| Year | Y     | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ |
|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 2010 | 849.8 | 825   | 68    | 45.7  | 8.55  | 1     | 67    | 2479  | 4009  |
| 2011 | 1783.6| 1028  | 65    | 45.3  | 18.9  | 7     | 60    | 2815  | 4760  |
| 2012 | 2490.5| 1235  | 57    | 45.4  | 10.1  | 2.6   | 57    | 2915  | 5098  |
| 2013 | 3042.6| 1465  | 54    | 45.8  | 16.3  | 2     | 63    | 3076  | 5683  |
| 2014 | 3696.4| 1722  | 49    | 45.4  | 17.4  | 4     | 65    | 3426  | 6159  |
| 2015 | 4036.7| 2000  | 45    | 46.1  | 17.7  | 3.1   | 63    | 3940  | 6845  |
| 2016 | 4792.2| 2205  | 41    | 45.7  | 20.1  | 3.5   | 68    | 4328  | 7019  |
| 2017 | 6335.7| 2400  | 41    | 46.3  | 19.7  | 3.3   | 64    | 4298  | 7721  |
| 2018 | 6987.5| 2550  | 38    | 46.8  | 20.9  | 4.6   | 61    | 4836  | 8219  |
| 2019 | 7498.3| 2693  | 34    | 46.5  | 18.9  | 5.4   | 66    | 5560  | 8734  |
| 2020 | 8343  | 2827  | 30    | 46.7  | 14.8  | 5     | 70    | 5917  | 9248  |

The data are all from China National Knowledge Infrastructure Statistical Database.

After normalizing the data in Table 1, the principal component analysis was performed by SPSS software to screen the elements that have great influence on the output variables. Firstly, the data were tested by KMO and Bartlett. The results are shown in Table 3.
Table 3. KMO and Bartlett test results.

|               |       |
|---------------|-------|
| KMO           | 0.684 |
| Approximate chi-square | 97.471 |
| Bartlett sphericity test | df 28 |
|               | p 0.000 |

From the above table can be seen: KMO is 0.684, greater than 0.6, and data through the Bartlett sphericity test \((p < 0.05)\), indicating that the research data is suitable for principal component analysis.

Then the influencing factors were analyzed by principal component analysis. The results are shown in Table 4.

Table 4. Variance explanation rate table.

| number | latent root | Variance explanation rate % | Cumulative % | latent root | Variance explanation rate % | Cumulative % |
|--------|-------------|------------------------------|--------------|-------------|------------------------------|--------------|
| 1      | 5.374       | 67.176                       | 67.176       | 5.374       | 67.176                       | 67.176       |
| 2      | 1.382       | 17.274                       | 84.450       | 1.382       | 17.274                       | 84.450       |
| 3      | 0.613       | 7.662                        | 92.112       | -           | -                            | -            |
| 4      | 0.383       | 4.786                        | 96.898       | -           | -                            | -            |
| 5      | 0.204       | 2.552                        | 99.450       | -           | -                            | -            |
| 6      | 0.033       | 0.415                        | 99.865       | -           | -                            | -            |
| 7      | 0.007       | 0.090                        | 99.955       | -           | -                            | -            |
| 8      | 0.004       | 0.045                        | 100.000      | -           | -                            | -            |

It can be seen from Table 3 that a total of two principal components are extracted from the principal component analysis, and the characteristic root values are greater than 1. The variance interpretation rates of the two principal components, namely, the annual training times of the enterprise management and the average age of the management, are 67.176 %, 17.274 %, and the cumulative variance interpretation rate is 84.450 %, that is, the two principal components can cover 84.450 % of the information of all data, which is representative. Results the load diagram is shown in Figure 2.
Thus select the number of annual training of enterprise management and the average age of management is $x_1, x_2$, the average annual total income of the enterprise is $y$, multiple regression analysis, the results are shown in Figure 3.

Expression is

$$y = 0.405x_1 - 0.538x_2$$

The results of model 7 are used to predict the annual average input of enterprises from 2018 to 2020. The results are as follows.

| Year | Real data | Predicted data | R2  |
|------|-----------|----------------|-----|
| 2018 | 6987.5    | 7200           |     |
| 2019 | 7498.3    | 7631           | 0.980|
| 2020 | 8343      | 7983           |     |

Analysis of error and correlation, the results are shown in Figure 4.
Figure 4. Breakline chart for comparison of predicted and true values from 2018 to 2020.

3.2. Discussion
From the results of principal component analysis, it can be seen that among the eight influencing factors existing in the current human resources management, the annual training times of enterprise management and the average age of management account for the largest proportion, and the weights in the extracted two principal components are 14.02 % and 13.35 %, respectively. It is proved that the enterprise human resources management is closely related to the advanced science of management concept and acceptance of advanced ideas. When the multiple regression model based on the two variables is applied to the annual income forecast of enterprises from 2018 to 2020, the annual income of enterprises is positively correlated with the number of management learning and negatively correlated with the average age. It is suggested that enterprises should increase the learning of scientific management concept of management personnel, attach importance to young talents, appropriately add young talents to the management, and reform the management system in real time, so as to improve the efficiency of human resource management and maximize the effect of management strategy. The correlation coefficient between the results of this model and the real value is 0.980, which is highly linear. It proves that the linear model established in this experiment is true and effective.

4. Conclusions
This paper constructs a human resource management evaluation model based on multiple regression from the perspective of linear system, so as to analyze human resource management strategies. Through experimental analysis, it is found that the linear multiple regression model is vulnerable to extreme variables and needs to establish long-term reports. Though the operation of the detection mechanism is complex, it is easy to understand and implement. In summary, through the linear method to deal with the collected data, we can carry on the comprehensive evaluation of human resources management, which is conducive to enterprise managers based on the company’s historical performance data evaluation of human resources management science, and has very important practical significance.
References

[1] Xie Lin, Du Gang. The evaluation index system and analysis model of human resource flexible management [J]. Journal of Xi'an University of Electronic Science and Technology (Social Science Edition), 2006 (02): 103-107.

[2] Zhou Liqun, Xia Liangke. Measurement and comparison of regional economic integration: Evidence from Beijing-Tianjin-Hebei, Yangtze River Delta and Pearl River Delta [J]. Jianghai Journal, 2010 (04): 81-87.

[3] Spatial econometric empirical analysis of the effect of heterogeneous human capital on economic growth [J]. Economic science, 2012 (01): 39-50.

[4] Chen Jiahui. Research on the interactive performance of human capital and economic development in the Yangtze River Delta based on the coupling coordination model [J]. China Economic and Trade Journal (China), 2021 (02): 34-39.