Journal of Advances in Applied Mathematics, Vol. 3, No. 4, October 2018
https://dx.doi.org/10.22606/jaam.2018.34006

A Study of the Education Return Based on R Language and Regression Analysis

Xinfeng Li[1], Hao Deng[*2]

[1]College of Liberal Arts, University of Minnesota, Twin Cities, Minneapolis, The United States
[2]School of Mathematics, University of Edinburgh, Edinburgh, The United Kingdom
[*]Email: Li000027@umn.edu; 564398527@qq.com

Abstract. To study the return on education, R language and regression analysis were applied. First, the research background and research status at home and abroad were introduced. Then, the relevant theories and influencing factors of educational returns were analysed. Finally, taking Hunan Province as an example, using the data surveyed by the National Bureau of Statistics, a multiple linear regression model was established to study the educational returns of urban residents. The results showed that education, employment and the degree of regional development all affected educational returns. Therefore, the level of investment in education should be raised and the regional gap should be narrowed.

Keywords: R language; regression analysis; educational return

1 Introduction

All civilizations in human society need to be based on the development of education. In the 1960s, human capital theory was founded [1]. Human capital theory believes that human capital, like other types of capital, can improve quality and production levels through investment channels such as education, training, and health care. Education is the most important way to improve human capital [2]. Governments in many countries around the world believe that education can promote long-term sustainable and rapid economic development by improving the quality of national workers and increasing labour productivity. At the same time, since the beginning of education as an investment behaviour, the research on the issue of educational returns has gradually become a research hotspot for some people concerned with education issues [3]. Many scholars believe that investment in education can bring about returns in terms of economic growth and per capita income. The return on education is influenced by many factors [4]. These factors mainly include education level, region, age, gender and so on.

2 State of the Art

Since the 1950s, economists represented by Schultz, Becker, and Mincer have conducted pioneering research on human capital theory [5]. Later, the study of the relationship between education and income began to be favored by many scholars. Becker used the cost-benefit method to estimate the rate of return of different groups of universities and high school education in the United States in the first half of the 20th century. The results show that as people become more educated, the return on education is lower. The difference in education is the most important factor affecting people's income differences [6]. According to the theory of human capital, Mincer used the difference compensation model and the accounting equation model to derive the famous Mincer salary equation model. The form of the Mincer salary equation model is very simple, so it has been widely used. Willis added the influence of ability variables on education. The role of problems such as ability bias and self-selection in educational return is studied. Interpretation of the relationship between educational choices, personal abilities and income is explained by establishing two production functions for relatively different occupations. It found that individuals with high abilities had higher incomes after receiving higher education.

Domestic scholars' research on educational returns mainly focuses on the introduction of analytical methods and basic principles. Zhong Yining studied the changing trend of the educational return rate of

Chinese urban residents. At the same time, the standard Mincer's income equation model is compared with the improved Mincer's income equation model. The trend of rising educational returns estimated by the standard model is obvious. Ge Yuhao used the survey data of Chinese urban households in 2000 for analysis. Women's education returns are higher. Age is inversely related to the rate of return to education [7]. Gao Mengyu used the endogenous treatment effect model to analyze the micro data of 7,949 households in three cities in western China. The educational investment return rate of urban youth aged 20-35 is calculated. It is found that the return on education investment of this group of people is roughly 7%, and the return rate of women in higher education is high [8].

3 Methodology

3.1 Related Concepts and Characteristics of Educational Returns

The return on education can be divided into direct return and indirect return. The direct return of education refers to the monetary income brought about by receiving education. The indirect return of education refers to the benefits of education other than money, such as personal protection, living ability, and various technologies. The return on education corresponds to the investment in education. Educational investment refers to the study of investing certain funds in a certain period to acquire some skills. It is the sum of the three aspects of manpower, material resources and financial resources expressed in the form of currency. According to the main characteristics of investment, education investment can be divided into personal investment and social investment. Personal investment is the cost of an individual's education. Social investment refers to the expenses that the whole society pays for education. The main source is the government's financial allocation and the donation of public welfare provided by individuals or organizations. As far as the current situation is concerned, the government has the largest proportion of investment in education, but the level of government investment and education is negatively correlated. Indirect investment in education refers to the loss of other opportunity costs due to investment education.

The return on education refers to the increase in monetary income of the state, family, individual or society as a result of investing in educational activities. As a personal and social investment, the rewards of education can be divided into individual returns and social returns, depending on the subject of the benefit. The personal reward for education refers to the increase in income earned by an educated individual because of receiving education. Generally, people with higher education levels have higher incomes, good benefits, and a lower risk of unemployment. The social return of education refers to the benefits that society derives from investing in education and other inputs that contribute to education. Education has played a very important role in the development of human civilization and today's social economy. The well-educated science and technology workers are vital to the development of the social economy, which is one aspect of the return of education society. Educational returns can also be divided into monetary returns and non-monetary returns. The monetary return is the increase in monetary income obtained from investment in education. Non-monetary returns refer to the sense of satisfaction gained from receiving education, the improvement of personal qualities, and changes in the working environment.

3.2 General Theory and Method of Educational Return

The return on education is also called the income of education. The return on education is usually obtained by calculating the rate of return on education. The most commonly used estimation method is the Mincer income equation method. The general form of the Mincer income equation is:

$$InW = a_0 + a_1S + a_2EXP + a_3EXP^2 + \mu \tag{1}$$

In this equation, InW is the natural logarithm of wages. a_1 is the regression coefficient, which represents the incremental income earned for each additional year of education, that is, the rate of return on education. S indicates the working time of the individual, and the age of the individual can be subtracted from the age at the end of the education. EXP is the individual's work experience. μ is the random error term. It represents the impact of factors other than the above two factors on individual income. The Mincer income equation has been widely used because of its simple form. One of the biggest features of this equation is the assumption that the rate of return on education is a constant. The rate of return for education for every educated person is equal. This assumption indicates that each person receives the same education and the human capital obtained is homogeneous. Human capital has two attributes, quantity and quality. Through such an assumption, the difference in human capital is reduced to the difference in quantitative attributes. However, the return of one year of primary or secondary education and one year of higher education is different. Moreover, the quality provided by these two kinds of education is also different. Therefore, the theory that the return on education is a constant is insufficient.

The two basic assumptions in the Mincer income equation were improved in the subsequent study of the rate of return on education. Other factors that have an impact on wages are added to the equation. These factors can be gender, occupation, region, and other control variables. The expression is as follows:

$$InW = a_0 + a_1 S + a_2 EXP + a_3 EXP^2 + \sum_{i=1}^{n} \alpha_i X_i + \mu \tag{2}$$

In this equation, α_i and X_i are added. Other control variables are represented by X_i. α_i represents the coefficient corresponding to these control variables. Compared with the first equation, the improved equation calculates the educational return rate more accurately.

There are many factors that influence the rate of return on education and the estimated rate of return on education. It mainly reflects the choice of sample, personal ability and its characteristics, and the estimation and test method of educational return. If the sample is not properly selected, it will cause a great deviation in the rate of return on education. If the deviation of the sample is overcome, the equation for selecting the sample should be set in the calculated model. Personal abilities and their characteristics are difficult to measure. Usually, it is converted into the form of money. The result of the calculation is different due to the difference in the conversion method. To overcome the deviation caused by this factor, it is generally necessary to add a proxy variable to the model or use the instrumental variable method to estimate the method.

3.3 Establishment of Regression Model

There are many nonlinear forms in the economic model. The general nonlinear regression model can be expressed as equation (3):

$$y = f\left(x, b\right) + \varepsilon \tag{3}$$

In the equation, x is an observable and independent random variable. b is the parameter vector to be evaluated. y is an independent observation variable. Its mean depends on x and b. ε is a random error.

The multiple linear regression parameters are determined by the least square estimation.

$$Q = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \tag{4}$$

In the formula, the overall weight of each sample point is the same. However, to minimize the overall impact of the outliers and not waste any information on a sample point, different weights are assigned to different sample points.

The core problem of the weighted multivariate nonlinear regression model is the construction of weight. The idea is as follows. First, $\left|y_i - \hat{y}_i\right|$ is calculated according to the general multivariate nonlinear regression model. Then, $\left|y_i - \hat{y}_i\right|$ is sorted. The order statistics Z_1, Z_2,...,Z_n is obtained. For n samples, the distribution function of the random variables in this sample space is set.

$$\begin{gathered} F\left(z\right) = P\left(Z \le z\right) \\ F^{-1}\left(\tau\right) = \inf\left\{z : F\left(z\right) \ge \tau\right\} \end{gathered} \tag{5}$$

Z is the quantile of τ. Then, according to the distribution function of Z, the corresponding value of $F^{-1}(0.2)$ is calculated, and it is set to Z_i. According to the formula, the Z values of $F^{-1}(0.4)$, $F^{-1}(0.6)$, $F^{-1}(0.8)$ and $F^{-1}(1.0)$ can be calculated separately. They are set up to Z_j, Z_m, Z_n and Z_l respectively. The criteria for setting weights are shown in Table 1.

Table 1. The criteria for setting weights

Quantile	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Evaluation value of Z_i index	9	7	5	3	1

According to the definition, the evaluation value of Z_1~Z_i is 9. The evaluation value of Z_{i+1}~Z_j is 7. The evaluation value of the Z_{j+1}~Z_m is 5. The evaluation value of the Z_{m+1}~Z_n is 3. The evaluation value of the Z_{n+1}~Z_l is 1. The evaluation value of the corresponding index for each Z_i is β_i .

$$Q' = \sum_{i=1}^{n} \rho_i Z_i^2, \qquad \rho_i = \frac{\beta_i}{\sum_{i=1}^{n} \beta_i} \tag{6}$$

Finally, the regression coefficient β_0 is determined by using the minimum principle of Q'. $\beta_0 = \left[\beta_1, \beta_2, ..., \beta_n\right]$ is selected as initial iterations. Then, it returns to the initial step until the estimated value of the estimated value β_ω is calculated. When the maximum value of the absolute value of the two stepwise regression coefficients is less than the set standard error, the iteration is over.

Based on the principle of weighted nonlinear regression model, the weighted multivariate nonlinear regression of V18-V23 is carried out respectively. Multiple linear regression was used in V18, V22 and V23. The indexes of V19, V20 and V21 were carried out in cubic, and then multiple linear regression was carried out.

4 Results and Discussion

4.1 Variable Description and Data Source

The data comes from the special survey project of the urban residents of Hunan Province by the National Bureau of Statistics. The educational level, employment situation, and the impact of the region on the return on education were studied. According to the per capita of counties in Hunan Province, the study area is divided into three regions: economically developed, medium and underdeveloped. After investigation, urban residents in economically developed areas accounted for 50%, followed by urban residents in moderately economically developed areas accounted for 26%. Urban residents in economically underdeveloped areas are equivalent to moderately economically developed regions, accounting for 24%. Regional distribution of residents is shown in Figure 1.

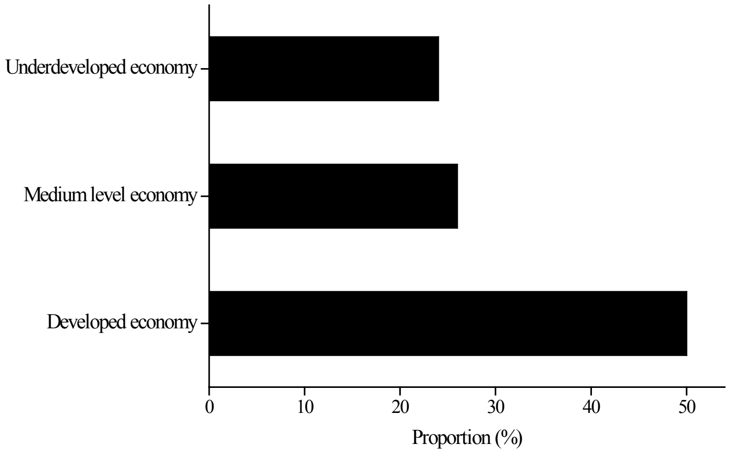

Figure 1. Regional distribution of residents

It can be seen from Figure 2 that the cultural level of the respondents is mostly high school, technical secondary school and junior high school and below, accounting for 34.4% and 34.2% respectively. Urban residents with college degrees account for 18.8%, and urban residents with bachelor degrees account for

11.6%. Only 1% of urban residents is postgraduate qualifications. The educational level of the respondents is shown in Figure 2.

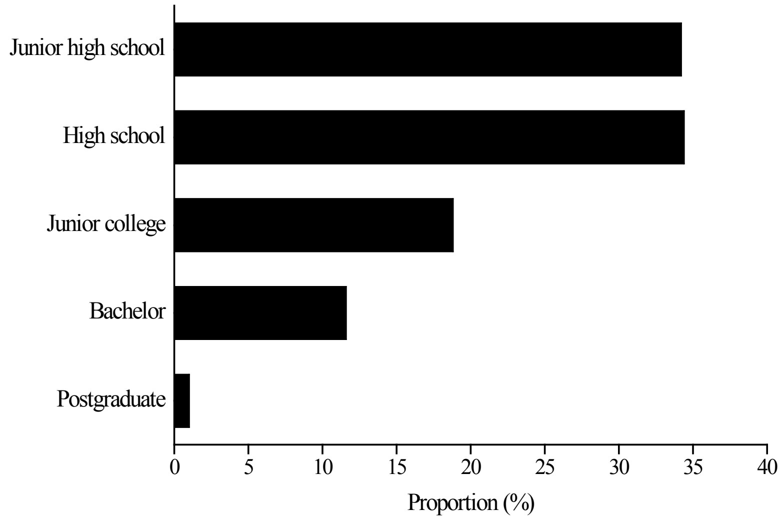

Figure 2. The educational level of the respondents

As can be seen from Figure 3, the employment of state-owned economic units accounted for 25.8%. The urban collective economic unit accounted for 2.3%. Other economic units accounted for 7.1%. Urban individual or private enterprise owners accounted for 10.1%. Urban individual or private enterprise employees accounted for 10.8%. Retired reemployed accounted for 1%. Other employed persons accounted for 11.5%. School students accounted for 2.5%. Retirees accounted for 28.8%.

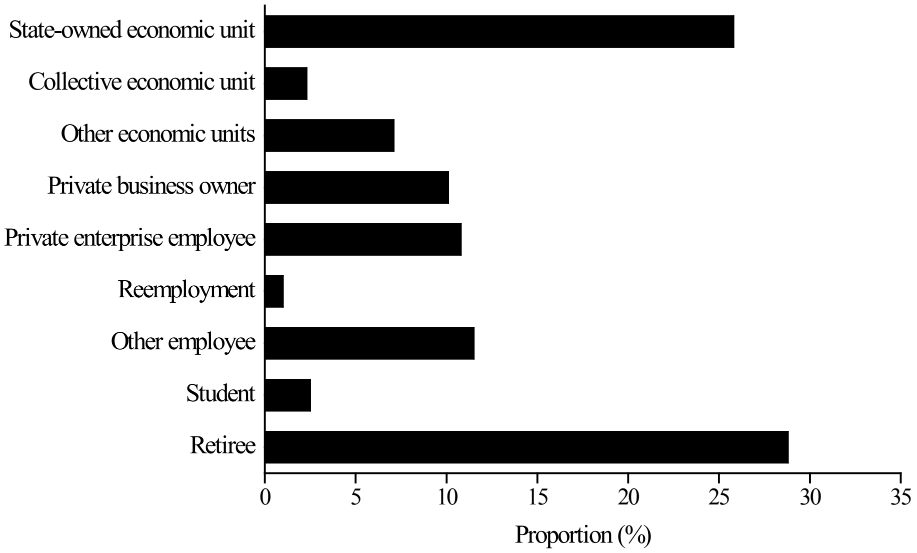

Figure 3. Employment status

4.2 Result Analysis

To study the employment methods of urban residents of different educational levels, the corresponding analysis method is applied to the two variables of education level and employment status. Regression analysis of employment situation and education level is shown in Table 2.

Table 2. Regression analysis of employment situation and education level

Dimension					Inertia ratio		Confidence singular value	
								Correlation
	Singular value	Inertia	Square	Sig.	Explanation	Accumulation	Standard deviation	2
1	0.498	0.248	-	-	0.954	0.954	0.016	0.103
2	0.088	0.008	-	-	0.030	0.984	0.017	-
3	0.051	0.003	-	-	0.010	0.994	-	-
4	0.041	0.002	-	-	0.006	1.000	-	-
Total	-	0.260	668.778	0.000[a]	1.000	1.000	-	-

As can be seen from Table 2, the inertia value of the variable in the first dimension is 0.248, which explains 95.4% of the total information. The inertia value in the second dimension is 0.008, which explains 0.03% of the total amount of information. The inertia value in the third dimension is 0.003, which explains 0.01% of the total amount of information. The inertia value in the fourth dimension is 0.002, which explains 0.006% of the total amount of information. Therefore, the two-dimensional graphics can better display the information between the two variables, and the first dimension is mainly observed.

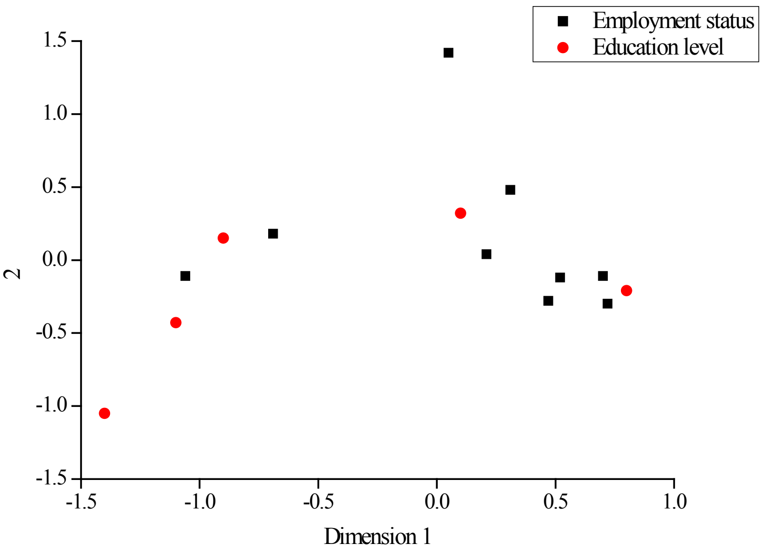

Figure 4. Regression analysis of employment situation and education level

From this, the following conclusions can be drawn: First, urban residents with tertiary education and above tend to work in state-owned economic units or other economic units. Second, urban residents with high education levels in high schools tend to work in private business units. Most of the urban residents with junior high school education are retirees, and most of them tend to start their own businesses.

To study the differences in the educational level of urban residents in different regions, the corresponding analysis method is applied to the degree of education and the region. Regression analysis of regional and educational level is shown in Table 3.

As can be seen from Table 3, the inertia value of the variable in the first dimension is 0.038, which explains 98.7% of the total information. The inertia value in the second dimension is 0.001, which explains 0.013% of the total amount of information. Therefore, the two-dimensional graphics can better display the information between the two variables, and the first dimension is mainly observed.

Table 3. Regression analysis of regional and educational level

Dimension					Inertia ratio		Confidence singular value	
								Correlation
	Singular value	Inertia	Square	Sig.	Explanation	Accumulation	Standard deviation	2
1	0.195	0.038	-	-	0.987	0.987	0.019	0.024
2	0.022	0.001	-	-	0.013	1.000	0.015	-
Total	-	0.038	98.749	0.000[a]	1.000	1.000	-	-

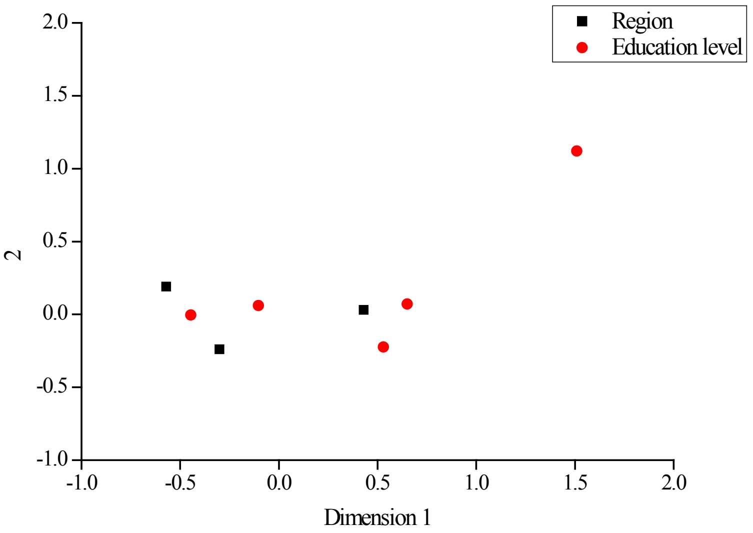

Figure 5. Regression analysis of regional and educational level

As can be seen from Table 3 and Figure 5, the educational level of urban residents in economically developed areas is higher than in the other two regions. This conclusion is consistent with the actual situation. Because in economically developed areas, the level of education, educational conditions and other educational facilities are also good.

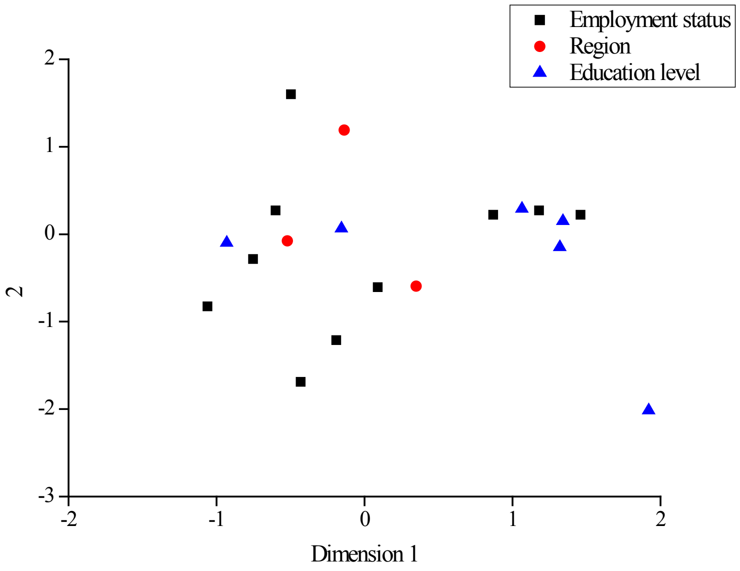

Figure 6. Regression analysis of region, education level and employment situation

To study the differences in the employment methods of urban residents in different regions and different educational levels, the corresponding analysis method is applied to the employment situation, education level and region. Regression analysis of region, education level and employment situation is shown in Figure 6.

As can be seen from Figure 6, urban residents in economically developed areas tend to work in collective economic units. Urban residents in the middle areas of economic development are mostly employees of private enterprises. Urban residents in underdeveloped areas are mostly retirees, urban individual or private enterprise owners. Urban residents with junior college and undergraduate degrees tend to work in state-owned economic units and other economic types. Most of the urban residents with junior high school education and below are retirees or retired reemployed people. The urban residents of high school education are mostly the owners of individual or private enterprises.

Based on this, the following findings can be obtained: First, in economically developed areas, urban residents with high education generally work in state-owned economic units and other economic types. Second, the qualifications of re-employed and retirees are generally junior high school and below. This result is in line with the actual situation. In the past, the level of education was relatively backward, and people's education was generally less educated. Therefore, it is normal for the re-employed and retirees to have low academic qualifications.

5 Conclusion

According to the National Bureau of Statistics survey data of urban residents in Hunan Province, the impact of employment, education, and regional development on education returns was studied. The following conclusions are drawn: The education level of urban residents in economically developed areas is relatively high. Most of them are undergraduates and tertiary institutions. The average education level of urban residents in middle economic development is junior high school and high school. The education level of urban residents in less developed areas is relatively low. Most of them are junior high school or below. Urban residents with higher education are generally in economically developed areas. They tend to work in state-owned economic units and other economic types. Less-educated urban residents are concentrated in economically underdeveloped areas. Urban residents in the middle areas of economic development are employees of individuals or private enterprises. With the development of the economy, the educational level of urban residents will change accordingly. In economically underdeveloped areas, the educational level and educational returns of urban residents are low.

References

1. Kuo, M. Y., & Shiu, J. L. (2016). A dynamic quantitative evaluation of higher education return: evidence from Taiwan education expansion. Journal of the Asia Pacific Economy, 21(2), 276-300.
2. Erdal, M. B., Amjad, A., Bodla, Q. Z., & Rubab, A. (2016). Going back to Pakistan for education? The interplay of return mobilities, education, and transnational living. Population, Space and Place, 22(8), 836-848.
3. Bhuller, M., Mogstad, M., & Salvanes, K. G. (2017). Life-cycle earnings, education premiums, and internal rates of return. Journal of Labor Economics, 35(4), 993-1030.
4. Wang, Q., Tang, L., & Li, H. (2015). Return migration of the highly skilled in higher education institutions: A Chinese university case. Population, Space and Place, 21(8), 771-787.
5. Assari, S. (2018). Blacks' Diminished Return of Education Attainment on Subjective Health; Mediating Effect of Income. Brain sciences, 8(9), 176.
6. Daly, A., Lewis, P., Corliss, M., & Heaslip, T. (2015). The private rate of return to a university degree in Australia. Australian Journal of Education, 59(1), 97-112.
7. Zimmerman, D. M., & House, P. (2016). Medication safety: Simulation education for new RNs promises an excellent return on investment. Nursing economics, 34(1), 49.
8. Lauder, H. (2015). Human capital theory, the power of transnational companies and a political response in relation to education and economic development. Compare: A Journal of Comparative and International Education, 45(3), 490-493.