Analysis of Factors Influencing the Off-Farm Employment Based on the Method of PLS

Ying Huang, Yizong Xu

Analysis of Factors Influencing the Off-farm Employment Based on the Method of PLS

Ying Huang [1], Yizong Xu [2]

[1, 2] School of Economics and Management, Beijing Forestry University
Beijing, 100083, P. R. China
huangying_1029@163.com

Abstract. With the widening income gap between the urban and rural areas, off-farm employment has undergone rapid development. The essay aims at analyzing the factors influencing the off-farm employment according to the statistics collected from Changle City in Fujian Province using the method of Partial Least-squares (PLS). Since current statistical methods such as the least squares, Logistics and the principle component analysis can not avoid the existence of the problems such as multiple correlation, single dependent variable and poorly explanatory information. However, these problems can be smoothed away by implementing the method of PLS. The results show that training experience of the rural labor force, education level, vocational skills, and health status are main factors influencing the off-farm employment while working area only has a slight impact. Some correspondent solutions and advice are available according to the analysis.

Keywords: Off-farm Employment, Partial Least-Squares (PLS), Fujian

1 Introduction

Large number of rural population, low per capita arable land, late start industry, low level of urbanization and the existence of urban-rural dual structure, are all basic conditions of current China. This condition has resulted in the oversupply of rural labor force for a long time. With the widening income gap between the urban and rural areas, the margin income of the agricultural labor is undergoing a further decrease. Quantities of rural labor migrate to the metropolis, which contributes greatly to the development of the off-farm employment. According to the figures from the National Bureau of Statistics, in the year 2008, the average wage income of the rural residents has reached to 1854 per capita. The increment of the wage income accounts for 41.5% of the whole year's total incremental net income. Based on the statistics, the increase of the wage income from the migrate workers contributes significantly to the family economy [1].

In the existing literature on rural off-farm employment of labor force, the current theory suffers several weaknesses: (1) studies pay more attention to income growth factors of the off-farm employment rather than job stability, which is an important indicator of the off-farm employment situation [2]. (2) The variables, which impact

off-farm employment in the rural labor force, still have multiple correlations. Implementing the simple least square method, the logistic [3] method and the principle component analysis method can hardly smooth away the limitation derived from the multiple correlation. The disadvantage of the multiple correlation lies in variables, which consist quantities of the duplicated information ,thus, exaggerating the status of the feature in the analysis system[4].(3)simply using the principle components analysis [5] can prevent us from interpretive understanding of the dependent variables, which may result in poor explanation based on the analysis of the independent variables.

PLS is widely used in establishing the statistic correlation between the independent variables and the dependent variables. In the modeling process, PLS regression analysis aggregates the advantages of principal component analysis, canonical correlation analysis and multiple linear regression method. When processing the independent variables which have multiple correlations, PLS could do better in modeling [6]. PLS makes component t_1 of the independent variables and component u_1 of the dependent variables contain the mutant information from the data sheet as much as possible. At the same time, t_1 and u_1 can achieve the greatest degree of correlation, which means t_1 and u_1 have great explanatory power [7].

2 Sources and the Sample Characterization

2.1 Figures

The data used in the essay are all from the survey of Wenwusha, Fujian. The survey is conducted by an academic group. The sample size is 20, which is randomly collected and analyzed in the line with the fundamental principles of the statistics. Being one of the largest provinces which have great labor export, Fujian has a large number of the migrate workers leaving their homes for off-farm work every year. In the year 1978, the 1st industry account for 75.1% of population of employment structure of Fujian Province. However, in the year 2007, the percentage has decreased to 32.7%. The gaps between the 1st and the 2nd industry, or the 3rd industry widened at first, then shortened afterwards [8]. The trend has manifested that the decrease of the agricultural labor force make the proportion of the off-farm employment in the rural area increase.

2.2 Sample characterization

This paper selects income, position, job stability as the indicators of evaluating the off-farm employment situation of rural labor force. The analysis involves 5 aspects: working areas, training experiences, education level, vocational skills and health status, which are all used in investigating the impact on the off-farm employment of the rural area.

Table 1. variable

	The Name of the Variables	Abbr.
Dependent variables	Income(RMB/year)	Y_1
	Position(1=the most senior managers;2=the junior managers; 3=the preliminary managers;4=general staff)	Y_2
	Job Stability(The average frequency of job changing per year)	Y_3
Independent variables	Education level(0=illiterate; 1=primary; 2=junior; 3=senior high/secondary ; 4=undergraduate/specialist ; 5=graduate)	X_1
	Health Status(1=good;2=median;3=poor)	X_2
	Working Area(1=village; 2=town but non-village; 3=county but non-town; 4=province but non-town; 5=outer province; 6=foreign country)	X_3
	Vocational Skills(0=none;1=preliminary;2=median;3=advanced)	X_4
	Training Experience(1=participated;0=never participated)	X_5

1) rural labor force is the main form of off-farm employment

According to the survey, migrant workers in the region accounts for a larger proportion of 65%. Among the migrant workers, 25% of the rural labor force has chosen to work in other provinces, 20% are the county workers and 10% are the town workers. The forms of the migrant workers are predominant agricultural work, waging in spare time, waging or doing business oriented work, off-farm work, etc.

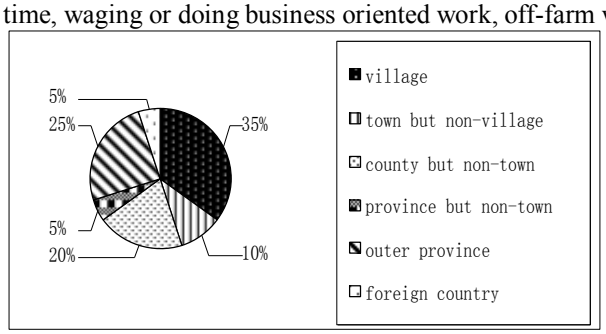

Fig. 1. Working area

2) Low education level of the rural labor force

In this region, the education situation of the rural labor force is not optimistic. The proportion of the rural labor force with education higher than the secondary level is only 5%. Few of the workers have postgraduate diploma or above. The workers with high school or secondary level of education account for the most, which is 30%, in this region. The number of workers dropping out of school after compulsory education is up to 20%. Only-primary school diploma holders account for 20% in the off-farm labor force of the rural area. Illiterate workers account for approximately

15%. The figures demonstrate that the rural labor force is generally in low quality, which posits impediment for their growth in the future.

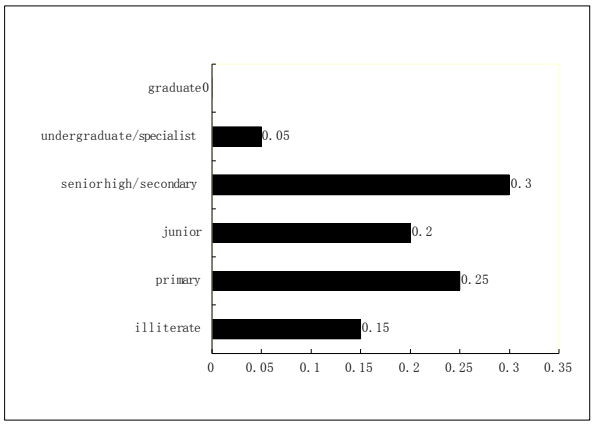

Fig. 2. the hierarchy of the vocational skills

3 The Analysis of PSL Regression

The basic idea of the PSL regression modeling is extracting t_1 and u_1 from the independent variables X(i) and dependent variables Y(i) respectively. The t_1 and the u_1 should contain the mutant information as much as possible from the variable sheet, that is. On the other way, it is required that the component t_1 has the greatest explanatory power to u_1. Based on the idea of the canonical correlation analysis, t_1 and u_1 should reach the maximum degree of the correlation, that is. Accordingly, in the PSL regression, the covariance between t_1 and u_1 should reach to the maximum value, that is:

$$Cov(t_1, u_1) = \sqrt{Var(t_1)Var(u_1)}r(t_1, u_1) \to \max \tag{1}$$

3.1 Accuracy analysis based on the PLS regression

In order to evaluate the predictive ability of the fitting equation, first, we should calculate the cross validation. For all of the dependent variables Y, the cross validation of component t is defined as:

$$Q_h^2 = 1 - \frac{S_{PRESS,\,hk}}{S_{SS,\,h-1}} \tag{2}$$

In the equation, represents the predictive error of the sum of squares, represents the sum of squares. When

$$Q_h^2 \geq (1-0.95)^2 = 0.0975 \quad (3)$$

the marginal contribution of t_h is significant.

Components:

A	R2X	R2X(cum)	Eigenvalues	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Significa
0	Cent.			Cent.					
1	0.547	0.547	2.74	0.397	0.397	0.316	0.05	0.316	R1
2	0.16	0.707	0.798	0.0859	0.483	0.0342	0.05	0.34	NS
3	0.108	0.815	0.542	0.0183	0.501	-0.202	0.05	0.274	NS
4	0.0631	0.878	0.316	0.00779	0.509	-0.128	0.05	0.201	N4

Fig. 3. Accuracy analysis

According to the Fig. 3, RdX represents the explanatory power of t_h to X and RdY represents the explanatory power of t h to Y. Q2 represents the cross validation. Based on the analysis, if a simply extracted validating component t_1 can explain 39.7% of the dependent variable in the variables set Y and the information utilization rate of the X variables set reaches 54.7%, introducing new principal component t_1 will significantly improve predictive capability of the model.

3.2 The analysis of the correlation of the dependent variables and independent variables

First, Fig. 4 demonstrates the values of the t_1/ u_1. In this Figure, t_1 is the first PSL component of the explanatory variable group; u_1 is the PSL component of the explained variable group. t_1 and u_1 are in the clear line form, which mean, the two group of the variables have a strong correlation. It is legitimate to grant the model validation.

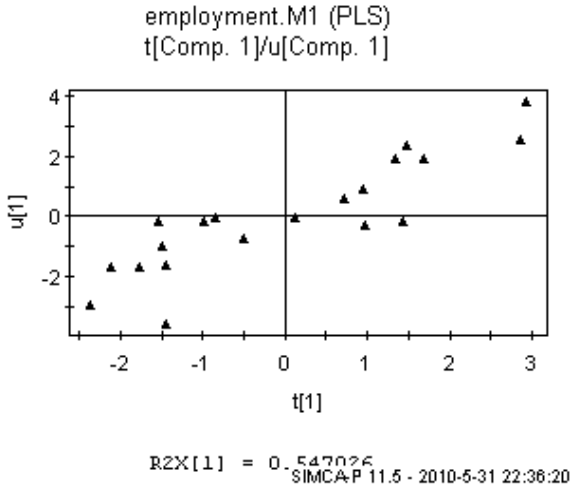

Fig. 4. Values of t_1/u_1

3.3 The effect when using independent variables to explain the dependent variables

Table 2. Values of VIP

Var ID (Primary)	M1.VIP[1]
X_2	1.16181
X_5	1.14651
X_1	1.03079
X_4	1.00695
X_3	0.509152

The effect that each independent variable X explains the dependent variable of the set Y can be evaluated by the important variable projection index VIP. When VIP>1, it shows that X has far more important effect on explaining the variable Y. Learning from Table 2, we can know that health status(X_2), education level (X_1), vocational skills(X_4), training experience(X_5) are all significant factors influencing the off-farm employment of the rural area. However, the working area(X_3) is inferior in influencing the off-farm employment of the rural area. Recently, government has paid more attention to the labor force of the rural area. With the development of the "Rural Labor Force Training Sunshine Project", the situations such as low level of education of the migrate workers, lacking necessary vocational skills have been ameliorated to some extent, making them more competitive in the process of the urbanization.

3.4 The discovery of the specific points

Generally speaking, since sample points which contribute excessively to the principle components can produce deviation when analyzing, we are trying to avoid the existence of such sample points. Therefore, we can measure the cumulative contribution rate that sample point i have to the components t_1, t_2……t_n.

$$T_i^2 = \frac{1}{n-1}\sum_{h=1}^{m}\frac{t_{hi}^2}{s_h^2} \tag{4}$$

SIMCA-P software use the Tracy statistics

$$\frac{n^2(n-m)}{m(n^2-1)}T_i^2 \sim F(m, n-m) \tag{5}$$

In the equation, n represents the number of the sample points, m represents the number of components used in the regression equation. When

$$T_i^2 \geq \frac{n^2(n-m)}{m(n^2-1)} F_{0.05}(m, n-m) \tag{6}$$

it can be confirmed that on the 95% test level, sample point i makes excessive contribution to the components t_1, t_2……t_n. Point i can be defined as the specific point ,which can result in deviation when analyzing. According to the Fig. 5, we can see that all the points are in the circumference of the ellipse. No specific point exists.

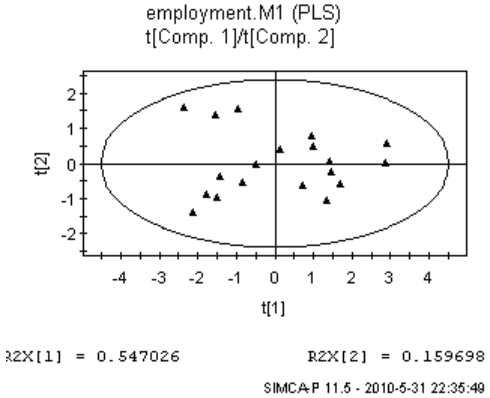

Fig. 5. T^2ellipse

3.5 Quality of the data reconstruction

When using the components t_1, t_2……t_n to establish the PSL regression model, because of omitting some original information, the difference of the fitted value and the actual value is too large, which makes it difficult to reconstruct the fitting equation. Under this scenario, we can measure the reconstruction quality of the sample points. According to this method, the distance of the sample points in the X space is:

$$s_i = DModX_i = \sqrt{\frac{e_{ij}^2}{p-m}} \times \sqrt{\frac{n}{n-m-1}} \tag{7}$$

In this equation, e_{ij}^2 represents the square of the difference of the fitted value and the actual value of the sample points. n presents the number of the sample. p represents the number of the independent variables. M represents the number of the

components in the regression equation. The average distance of the model in the set of sample points is defined as:

$$s_X = \frac{1}{n}\sum_{i=1}^{n} s_i^2 \tag{8}$$

The standard distance is:

$$(DModX, N)_i = \frac{s_i}{s_X} \tag{9}$$

3.6 The predictive result of the model

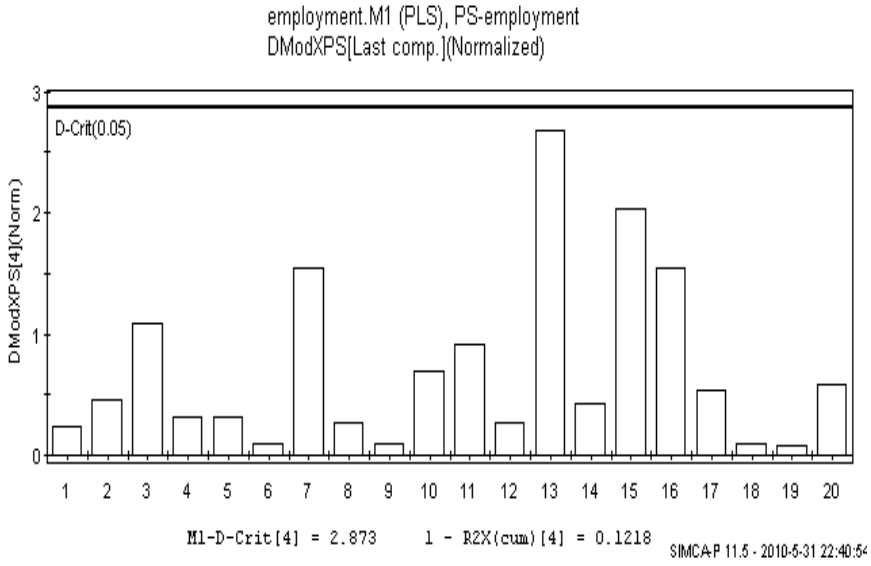

Fig. 6. The standard distance of the sample points

Demonstrated from the Fig. 6, all the values of distance vary from 0 to 2, which mean the reconstruction quality of the sample points is uniform.

4 Suggestions and solutions

According to the study, the labor force training, education, vocational skills, and health status are main factors influencing the off-farm employment while working area only has a slight impact.

In order to solve the problems mentioned above, the government should increase the training of the rural labor force and the training should focus on improving the professional skills of rural labor force. In order to improve the quality of the rural labor force, we should as well strengthen the rural education, implement variety of effective education forms and continue promoting the "Rural Labor Force Training Sunshine Project". Only thus can we truly ameliorate the situation of the off-farm employment in the rural area.

References

1. Du Yang, Piao Zhishui. Labor migration income and poverty [J]. China's rural observation, 2003，(5)：2—9．
2. Ren Guoqiang, Xue Shougang. Training and employment income growth of Chinese agricultural mechanization, impact study. 2009(06)
3. Xin Ling, Jiang Heping. Rural labor non-farm payrolls factors analysis_ Based on a rural labor force of 1006 Sichuan. Agricultural technology economy.2009(06)
4. Wang Huiwen, partial least-square regression method and its application. Beijing: defense industry press, 1999.4
5. Chen Xian, Huang Jianbai The factors affecting the migrant workers principal component analysis, 2009(18)
6. Ren Ruoen, Wang Huiwen, multivariate statistical data analysis. - theory, method and examples. Beijing: defense industry press, 2009:149
7. Jiang Yong, fujian industry structure and employment structure of correlation analysis, 2009. Technology (9),
8. Wang Huiwen, Wu Bin, Meng Jie. Partial Least-squares regression of linear and nonlinear method. Beijing: defense industry press, 2006.9