International Journal of Smart Home
Vol. 10, No. 11, (2016), pp. 219-228
http://dx.doi.org/10.14257/ijsh.2016.10.11.19

The Research of Multiple Regression Analysis in Rural-Urban Income Disparity

Jian Li[1,2] and Xiangyu Guo[1*]

[1]*Northeast Agricultural University*
[2]*First Affiliated Hospital of Heilongjiang University of Chinese Medicine*
lijian_hucm@163.com

Abstract

The multiple linear regression model contains more than one predictor variable and it shows the relationship among multiple variables. In the existing research field of rural-urban income disparity, the method of multiple regression analysis is mainly employed. But the linear relationship among variables is estimated mainly depending on principal component analysis. Principal component analysis is used to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The principal component analysis is widely used for feature extraction to reveal the most main factors from the multiple aspects. A multiply linear regression model integrating principal components analysis is proposed to address on the income gap between the city and country. The influential factors are given and the analysis results are discussed in this paper. The experimental results on income data from 1990 to 2013 show that the proposed method is effective in predicting the income ratio and analyzing the influential factors.

Keywords: *Multiple Linear Regression Model, Principal Component Analysis, Rural-urban Income Disparity*

1. Introduction

Linear regression is an approach for modeling the relationship between a dependent variable y and one or more independent variables) denoted X in statistic. For more than one explanatory variable, the process is called multiple linear regression [1].

The linear predicting functions are used to model the relationship among variables and the parameters of unknown model will be estimated from the data. The estimated models are generally called linear models [2]. Most commonly, the conditional mean of y given the value of X is assumed to be an function of X. Like other forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis.

There are many practical uses in linear regression. Most applications fall into one of the following two broad categories: (1) If the goal is prediction, we can use linear regression to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y. (2) Given a variable y and a number of variables X_1, ..., X_p that may be related to y, linear regression analysis can be applied to quantify the strength of the relationship between y and the X_j, to assess which X_j may have no relationship with y at all, and to identify which subsets of the X_j contain redundant information about y [3]. This paper focus on the two application of linear regression to: (1) predict the income ratio of rural-

*Corresponding Author

ISSN: 1975-4094 IJSH

urban income disparity, and (2) reveal the correlated influential factors of rural-urban income disparity.

This paper chooses nine features including urbanization level, rate of supporting agriculture in finance, growth rate of GDP *et al*. By using the principal component analysis, we construct the regression model to predict the income ratio. And the correlated influential factors are also analyzed by using the proposed method.

This paper is organized as followed. In Section 2, we discuss the multiple linear regression model and the principal component analysis is introduced. In Section 3, the regression model integrating principal component analysis is constructed. In Section 4, we give an example analysis by using rural-urban income data from 1990 to 2013. And our method is concluded in Section 5.

2. Multiple Regression Analysis Integrating Principal Component Analysis

2.1. Multiple Regression Analysis

A regression model relates Y to a function of X and β.

$$Y \approx f(X, \beta) \tag{1}$$

The approximation is usually formalized as

$$E(Y \mid X) = f(X, \beta) \tag{2}$$

If we have the knowledge about the problem domain, then the form of this function f can be estimated based on knowledge to model the relationship between Y and X. If no such knowledge is available, the function f can be estimate relying on the data. When we have a training dataset, we can estimate the function f to carry out regression analysis. Then the form of f can be specified.

Assume now that the vector of unknown parameters β is of length k. The following information about the dependent variable Y must be provided in order to perform a regression analysis:

The most common situation is where N > k data points are observed. In this case, there is enough information in the data to estimate a unique value for β that best fits the data in some sense, and the regression model when applied to the data can be viewed as an overdetermined system in β. In this case, the regression analysis provides the tools for finding a solution for unknown parameters β that will, for example, minimize the distance between the measured and predicted values of the dependent variable Y (also known as method of least squares). Under certain statistical assumptions, the regression analysis uses the surplus of information to provide statistical information about the unknown parameters β and predicted values of the dependent variable Y.

The model for regression is one that involves a linear combination of the input variables:

$$f(x, w) = w_0 + w_1 x_1 + \ldots + w_n x_n \tag{3}$$

where $x = (x_1, \ldots, x_n)^T$. The key attributes of this model is the estimation of parameters $w_0, \ldots, w_n$. It is a linear function of the input variables x_i, and this imposes significant limitations on the model.

We can consider the linear combinations of fixed nonlinear functions of the input variables by using the following form:

$$f(x, w) = w_0 + w_1\varphi_1(x) + \ldots + w_n\varphi_n(x) \quad (4)$$

where $\varphi_j(x)$ are known as basis functions. By denoting the maximum value of the index j by m, the total number of parameters in this model will be m. For the sake of allowing parameter w_0 as any fixed offset in the data, the linear regression function can be defined as the folloing form:

$$f(x, w) = \sum_{j=0}^{m} w_j\varphi_j(x) \quad (5)$$

where $w = (w_0, \ldots, w_m)^T$ and $\varphi = (\varphi_0, \ldots, \varphi_m)^T$. We will apply some form of fixed pre-processing, or feature extraction, to the original data variables. If the original variables comprise the vector x, then the features can be expressed in terms of the basis functions $\{\varphi_j(x)\}$.

2.2. Principal Component Analysis

Principal component analysis (PCA) is an analogue of the principal axis theorem in mechanics invented in 1901 by Karl Pearson [4]. Since it uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components, we commonly called it a statistical procedure.

Principal component analysis is mostly used as a tool in exploratory data analysis and for making predictive models. It can be done by eigenvalue decomposition of a correlation matrix or singular value decomposition of a data matrix [5]. PCA is one of the common tools of for multivariate analyses. Often, it is regarded as revealing the internal structure of the data in a way that best explains the variance in the data.

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on [6].

Consider a data matrix X, where each of the n rows represents a different repetition of the experiment, and each of the p columns gives a particular kind of feature.

Mathematically, the transformation is defined by a set of p-dimensional vectors of weights or loadings $w_{(k)}=(w_1,...,w_p\}_{(k)}$ that map each row vector $X_{(i)}$ of X to a new vector of principal component scores $t_{(i)} = (t_1, ..., t_k)_{(i)}$, given by

$$t_{k(i)} = w_{(k)} \cdot x_{(i)} \quad (6)$$

The Algorithm 1 is a detailed description of PCA using the covariance method as opposed to the correlation method [7].

3. Predicting Rural-Urban Income Disparity Ratio by Integrating Multiple Linear Regression Model and Principal Component Analysis

3.1. Prediction Model

By using principal component analysis, the number of principal components will be less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. Getting the vectors of weights, we can built the multiple linear regression model by

$$y=W.X+\beta \quad (7)$$

where y is the dependent variable, $W=(w_1,w_2,...,w_n)$ is the weight vector, $X=(x_1,x_2,...,x_n)$ is feature variable vector and β is an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressor.

3.2. Feature Variables

For predicting the rural-urban income disparity ratio, we choose nine influencing factors as the feature variables, which described in Table 1.

Table 1. Influencing Factors in Rural-Urban Income Disparity Ratio

Feature	Features Name	Meaning
X_1	urbanization level	urban population/the total number of people
X_2	rate of supporting agriculture in finance	fiscal expenditure used in rural areas/the total fiscal expenditure
X_3	growth rate of GDP	GDP value-added/ prior year's GDP
X_4	economic degree of opening-up	gross import and export/GDP of the whole country
X_5	the dual economic structure coefficient	radio of the second, third industrial output and Non-agricultural population/ radio of the first industrial output and agricultural population
X_6	Registered Unemployment Rate in Town	Registered Unemployment Rate in Town
X_7	Employment Structure	employment personnel of the second, third industrial/the total employment
X_8	Human capital differences between urban and rural	average annual per capital expenditure on education and culture of urban households/average annual per capital expenditure on education and culture of rural families
X_9	Urban and rural financial scale	total loan in urban areas/total loan in rural areas

Algorithm 1. Computing PCA Using the Covariance Method

Input: n data vectors $X_1,...,X_n$ with each X_i representing a single grouped observation of the *p* variables.
Output: Principal component scores T

1 the Kosambi-Karhunen–Loève transform (KLT) of matrix
$X: Y = \mathrm{KLT}\{X\}$

2 For each dimension $j = 1, ..., p$
Calculate the empirical mean

$$\mu[j] = \frac{1}{n}\sum_{i=1}^{n} X[i, j]$$

3 Calculate the deviations from the mean
$$B = X - hu^T$$

4 Find the covariance matrix:
$$C = \frac{1}{n-1} B^* \cdot B$$

5 Find the eigenvectors and eigenvalues of the covariance matrix
$$V^{-1}CV = D$$

6 Rearrange the eigenvectors and eigenvalues

7 Compute the cumulative energy content for each eigenvector

8 for j=1..k
$$g[j] = \sum_{k=1}^{j} D[k, k]$$

9 for j = 1..p
$$W[k, l] = V[k, l]$$

10 Select a subset of the eigenvectors as basis vectors
for k = 1, . . . , p l = 1, . . . , L $\frac{g[L]}{g[p]} \geq 0.9$

11 Convert the source data to z-scores (optional)
$$s = \{s[j]\} = \{\sqrt{C[j, j]}\}$$

12 Calculate the $n \times p$ z-score matrix
$$Z = \frac{B}{h \cdot s^T}$$

13 Project the z-scores of the data onto the new basis
$$T = Z \cdot W = \mathrm{KLT}\{X\}$$

4. Example Analysis

We take influencing factors of the rural-urban income disparity as the example to analyze the disparity ratio in this Section.

4.1. Experimental Dataset

In order to further understand the effects of above influencing factors in rural-urban income disparity, we collect the following data of X province from 1990 to 2013. We show the values of influence factors in each year in Table 2. Based on data availability and other considerations, the survey indicators mainly affect on rural-urban income disparity.

Table 2. Experimental Dataset: Values of Influence Factors in Each Year

Year	X1	X2	X3	X4	X5	X6	X7	X8	X9
1990	0.2641	0.531	0.0985	0.2962	10.3829	0.025	0.399	3.4994	2.8661
1991	0.2694	0.519	0.1662	0.33	11.5859	0.023	0.403	3.3411	2.9891
1992	0.2746	0.502	0.2362	0.3369	13.0839	0.023	0.415	3.3687	3.1012
1993	0.2799	0.42	0.3124	0.3173	14.3886	0.026	0.436	3.3306	3.2513
1994	0.2851	0.406	0.3641	0.4206	13.5671	0.028	0.457	3.3393	3.4682
1995	0.2904	0.4	0.2615	0.3844	13.2396	0.029	0.478	3.0553	3.7876
1996	0.3048	0.396	0.1708	0.3372	12.8986	0.03	0.495	2.8307	4.0219
1997	0.3191	0.39	0.1098	0.3395	13.8606	0.031	0.501	3.0259	4.0678
1998	0.3335	0.389	0.0687	0.3163	14.3633	0.031	0.502	3.1327	4.1152
1999	0.3478	0.387	0.0625	0.3315	15.1856	0.031	0.499	3.3687	4.3152
2000	0.3622	0.382	0.1063	0.3936	16.7615	0.031	0.5	3.3624	4.2069
2001	0.3766	0.374	0.1052	0.3825	17.2114	0.036	0.5	3.5818	4.3368
2002	0.3909	0.358	0.0973	0.4246	17.1633	0.04	0.5	4.2902	4.6414
2003	0.4053	0.35	0.1286	0.5161	17.1262	0.043	0.509	3.9646	4.7002
2004	0.4176	0.333	0.1768	0.5945	15.0217	0.042	0.531	4.1708	4.7572
2005	0.4299	0.329	0.1567	0.629	16.4009	0.042	0.552	3.7142	4.7323
2006	0.4434	0.325	0.1709	0.6477	17.5774	0.041	0.574	3.9427	4.6097
2007	0.4589	0.323	0.2314	0.6226	17.6152	0.04	0.592	4.3485	4.2204
2008	0.4699	0.303	0.1818	0.568	17.3882	0.042	0.604	4.3184	4.2028
2009	0.4834	0.328	0.0912	0.4359	17.8822	0.043	0.619	4.3245	4.2919
2010	0.4995	0.319	0.1831	0.4933	18.094	0.041	0.633	4.4384	4.1337
2011	0.5127	0.318	0.184	0.4883	17.8505	0.041	0.652	4.6719	3.8623

2012	0.5257	0.319	0.1033	0.4571	17.3806	0.041	0.664	4.5646	6.3157
2013	0.5373	0.323	0.1009	0.439	17.1708	0.0405	0.686	4.7211	3.4199

4.2. Principal Component Analysis on Experimental Dataset

For the correlation of influencing factors in rural-urban income disparity ratio, we using principal component analysis to analyze the experimental dataset. Table 3, shows the communalities of influencing factors.

Table 3. Communalities

Feature	Initial	Extraction
X_1	1.000	.985
X_2	1.000	.906
X_3	1.000	.945
X_4	1.000	.830
X_5	1.000	.847
X_6	1.000	.941
X_7	1.000	.897
X_8	1.000	.891
X_9	1.000	.915

According to the information extracted from the Table 3, which described the variables communalities, the information extracted from all of the original indicators are at about 90%. In other words, the comprehensive factors can better explain each of the original influenced factors.

The Total Variance Explained is described in Table 4.

Table 4. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.312	70.134	70.134	6.312	70.134	70.134
2	1.154	12.821	82.955	1.154	12.821	82.955
3	.692	7.692	90.647	.692	7.692	90.647
4	.346	3.841	94.488			
5	.228	2.532	97.020			
6	.177	1.967	98.987			
7	.073	.813	99.801			
8	.016	.177	99.978			
9	.002	.022	100.000			

According to Table 5, the total variance explained, it can be seen that we can extract three principal components from the policy factors of nine variables. And three principal components concentrated 90.647% original variable information. So it is ideal for the information content contained in the original variables.

The component matrix and the component score coefficient matrix are shown in Table 5 and Table 6.

Table 5. Component Matrix

	Component	
Feature	1	2
X_1	.967	-.023
X_2	-.935	-.080
X_3	-.265	.930
X_4	.769	.424
X_5	.920	.008
X_6	.968	-.007
X_7	.919	-.010
X_8	.847	.107
X_9	.706	-.302

Table 6. Component Score Coefficient Matrix

	Component	
Feature	1	2
X_1	.967	-.023
X_2	-.935	-.080
X_3	-.265	.930
X_4	.769	.424
X_5	.920	.008
X_6	.968	-.007
X_7	.919	-.010
X_8	.847	.107
X_9	.706	-.302

When analyzing principal component load coefficient described in Table 5, for the first principal component variables, the loads of nine influence factors of x2 and X3 is negative. For the second principal component variables, the loads of X1, X2, X6, X7 and x9 is also negative, and the third principal component variables, the load of the X3, X4, X5 and x9 is negative. According to the above results, this study and theoretical judgment are consistent. So it is feasible by using principal components on behalf of the relevant factors.

According to Table 6, the linear expression of standardized original variables can be estimated by principal component score coefficient matrix as follows:

$$F_1=0.153*2.512X_1-0.148*2.512X_2-0.042*2.512X_3+0.122*2.512X_4+0.146*2.512X_5+0.153*2.512X_6+0.146*2.512X_7+0.134*2.512X_8+0.112*2.512X_9$$

$$F_2=-0.020*1.074X_1-0.069*1.074X_2+0.806*1.074X_3+0.367*1.074X_4+0.007*1.074X_5-0.006*1.074X_6-0.009*1.074X_7+0.092*1.074X_8-0.262*1.074X_9$$

$$F_3=0.325*0.832X_1+0.232*0.832X_2-0.143*0.832X_3-0.3532*0.832X_4+0.011*0.832X_5-0.097*0.832X_6+0.334*0.832X_7+0.583*0.832X_8-0.824*0.832X_9$$

4.3. Predicting by Multiple Regression Analysis Model

By using principal component analysis, the model is summarized in Table 7. And the ANOVA is shown in Table 8.

Table 7. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.960	.921	.910	.1030901

Table 8. ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.490	3	.830	78.111	.000[b]

	Residual	.213	20	.011		
	Total	2.703	23			

Then take rural-urban income disparity ratio as dependent variable, and the above three principle components as independent variable, we get the linear regression model as follows:

$$Y=2.913+0.306F_1+0.116F_2-0.033F3 \quad (6)$$

Bringing the expression of principal component analysis into regression function, we can obtain the linear regression model as

$$Y=2.913+0.1062X_1-0.129X_2+0.072X_3+0.149X_4+0.113X_5+0.120X_6$$

$$+0.102X_7+0.098X_8+0.076X9$$

By the linear regression model we can conclude that the influenced factors X_1 and X_3 to X_9 is positive, so they have the positive effects on y. In other words, they play an expanded roles on income gap between urban and rural. But the weight of X_2 is negative, so it will reduce the urban-rural income gap.

5. Conclusion

In this paper, we proposed a model by using multiple linear regression analysis integrating principle component analysis. This model exploits the advantages of principle component analysis to estimate the importance of influence factors in urban-rural income gap. The weights of influence factors are learned by PCA, and combine the principle component into the linear regression function to prediction the urban-rural income gap ratio. Applying our method on the example dataset, the experimental results show that the proposed model can predict the all kinds of influence factors in urban-rural income gap.

References

[1] D. A. Freedman, "Statistical Models: Theory and Practice", Cambridge University Press, **(2009)**, pp. 26-29.
[2] A. C. Rencher, "Methods of multivariate analysis", John Wiley & Sons, **(2003)**.
[3] Y. Anzai, "Pattern Recognition & Machine Learning", Elsevier, **(2012)**.
[4] K. Pearson "LIII. On lines and planes of closest fit to systems of points in space", The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, **(1901)**, pp. 559-572.
[5] H. Abdi and L. J. Williams, "Principal component analysis", Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 4, **(2010)**, pp. 433-459.
[6] I. Jolliffe, "Principal component analysis", John Wiley & Sons, Ltd, **(2002)**.
[7] Engineering Statistics Handbook. Retrieved **(2015)** January 19.

Authors

Jian Li, Born in October 1982, Candidate Doctor, economist. His research interests include agricultural economic management, human resource management, *et. al*.

Xiangyu Guo, Born in July 1965, Doctor, Professor. His research interests include agricultural economic management, rural cooperative economy, *et. al.*