ISSN : 0974 - 7435 Volume 10 Issue 22

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(22), 2014 [13853-13860]

Unary linear regression method on principal component analysis

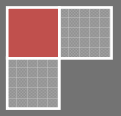
Xu Li-Li
School of Economics and Management, Henan Polytechnic University, Jiaozuo Henan 454000, (CHINA)

ABSTRACT

Only dependent variable is considered as random variables in commonly-used linear regression methods, so the regression results will be changed according with the coordinates selection. Enlightened by the method of principal component analysis (PCA), a new unary linear regression which is irrelevant to coordinates is proposed, which is the PCA based method. Compared with conventional least squares method, the new method possesses the advantages of lower deviation error and higher regression accuracy, which is verified by simulation cases and living examples. The simulation case and living example verified new method less system deviation and better regression accuracy than conventional. PCA is numerical solution, the advantage of low calculation amount makes it own a broad application prospect.

KEYWORDS

Unary linear regression; Least squares linear regression; Principal component analysis; Coordinate-independent.



INTRODUCTION

Linear regression analysis[1,2] is one of the most basic research method in Mathematical Statistics, which can be used to study the relationship between variables. In the field of socio-economics, even though relationship between many variables is not linear at macro level, it can still be linearized approximately at the micro level. In addition, sometimes through the preprocessing, such as logarithm of the variables, linear relationship between variables can be transformed to a linear relationship. Current main software for statistics analysis and numerical calculation is based on matrix operations. So it plays an important basic role to make a high precision linear regression analysis for variables.

UNARY LINEAR REGRESSION METHOD BASED ON PRINCIPAL COMPONENT ANALYSIS

Linear regression can be classified as unary, binary, and multiple diverse situations based on the number of independent variables and dependent variables. Among them, unary linear regression is the most basic problem, the derivation is as follows:

Assume that x, y conform to the linear relationship equation.

$$y = \beta_0 + \beta_1 x + \varepsilon \quad \textbf{(1)}$$

Here, β_i (i=0,1) is constant; ε is random error. Each variable is observed n times, the observation vector is:

$$X = (x_1, x_2, ..., x_n)' \quad \textbf{(2)}$$

$$Y = (y_1, y_2, ..., y_n)' \quad \textbf{(3)}$$

The above data is equal to the scatter set S, $S = \{(x_i, y_i) | i \in [1, n]\}$. Based on the observation data above, a unary linear regression line about variables x, y is[3]:

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0 \quad \textbf{(4)}$$

The matrix form of formula (1) is:

$$Y = (1, X)B + E \quad \textbf{(5)}$$

Here, $B = (\beta_0, \beta_1)', E = (\varepsilon_1, ..., \varepsilon_n)$.

The most commonly used solution method of unary linear regression is based on the least square linear regression: y is regarded as the dependent variable, x is regarded as independent variables, independent variable is not treated as random variable, only the dependent variable as random variable, the maximum likelihood estimation of parameter matrix B is[4]:

$$\hat{B} = (\hat{\beta}_0, \hat{\beta}_1)' = ((1, X)'(1, X))^{-1}(1, X)'Y \quad \textbf{(6)}$$

The result of least squares linear regression does not have the coordinate independence[5-7]. Coordinate-independence means that the orthogonal transformation of calculating coordinate system (translation and/or rotation) does not affect the result of the calculation. As shown in Figure 1, lines L and L' are the least squares results in the coordinate system for the same set of data in xOy and $x'O'y'$ respectively, and it is obvious that L and L' do not coincide.

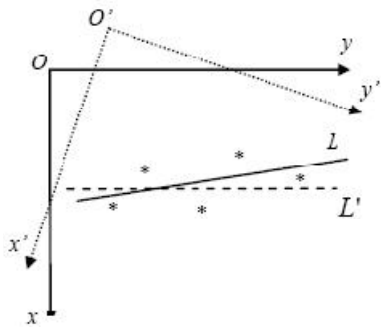


Figure 1 : Least square linear regressions in different coordinate systems

There are few 'pure' independent variables without randomness in socioeconomic variables. Due to the observation perspective, observation instruments, data definitions and different aggregation methods, observation data of the same economic phenomenon may have a big difference in form, but after some linear transformation, or even a simple coordinate transformation, it will show a clear equivalence between the data. Based on the above reasons, the same regression results of data sets with equivalence relations is a very natural requirement. So it is essential to develop a linear regression method with coordinate-independence.

Enlightened by the method of Principal Component Analysis (PCA), a new unary linear regression which is irrelevant to coordinates is proposed:

Definition 1: Variables x, y and their relationship are shown in formula (1)~(4), x, y are regarded as random variables, then PCA is carried out on vector set (X,Y), the slope of the first principal component corresponds to the vector will be obtained, which is supposed as $\hat{\beta}_1$. Assuming that the linear regression line pass through the barycenter $c(\overline{X},\overline{Y})$ of all the sample points, and its slope is $\hat{\beta}_1$, a linear regression equation (4) can be attained by using point slope method, x, y are variables, which are based on observation values (X,Y). The unary linear regression method above is called Principal Component Analysis Based Simple Linear Regression (PCABSLR), referring to principal component regression method[8,9].

The concrete solution process of principal component regression method is shown below:

Assume the covariance matrix of (X,Y) is:

$$\sum = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix} \tag{7}$$

Here,

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 \tag{8}$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^2 \tag{9}$$

$$\sigma_{xy} = \sigma_{yx} = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \overline{X})(y_i - \overline{Y})\right] \tag{10}$$

Calculate eigenvalues of Σ which corresponds to the first principal component：

$$\lambda_1 = \frac{1}{2}(\sigma_x^2 + \sigma_y^2 + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}^2}) \tag{11}$$

an eigenvector corresponding to λ_1 is:

$$\alpha_1 = \begin{pmatrix} 1 \\ \dfrac{\lambda_1 - \sigma_x^2}{\sigma_{xy}} \end{pmatrix} \tag{12}$$

The slope of α_1 is:

$$\hat{\beta}_1 = \frac{\lambda_1 - \sigma_x^2}{\sigma_{xy}} = \frac{-\sigma_x^2 + \sigma_y^2 + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}^2}}{2\sigma_{xy}} \tag{13}$$

Assumed that variable group regression line pass through barycentre $c\left(\overline{X},\overline{Y}\right)$ of each point in S, then,

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{14}$$

Here,

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{15}$$

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{16}$$

The results of principal component regression method is coordinate-independent[10], to prove it and to further reveal the nature of the principal component analysis, the following question is raised.

Example 1: S is a set of points in the plane P, $S = \left\{(x_i, y_i) \middle| i \in (1, n)\right\}$, find a straight line L in the plane, making the square of the distance from each point to line L in S is minimum.

Solution: assume L: $y = kx + b$ is a line in the plane P, the square of distance between each point in P and L is:

$$A = \sum_{i=1}^{n} \frac{(y_i - kx_i - b)^2}{1 + k^2} \tag{17}$$

Find the minimum point of $A(k,b)$, then,

$$\begin{cases} k = \dfrac{-\sigma_x^2 + \sigma_y^2 + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}{}^2}}{2\sigma_{xy}} \\ b = \overline{Y} - k\overline{X} \end{cases} \tag{18}$$

Compare the result of Example 1 with the result of principal component analysis in the previous section, it is obvious that $\hat{\beta}_1 = k$, $\hat{\beta}_0 = b$, which means that the two methods are equivalent, or principal component analysis is another form of least squares method. Regression line obtained by it is the line which has the minimum sum of square distance between the obtained line and every sample point. Based on the above analysis, it is obvious that the results of principal component regression method is coordinate-independent.

Simulation experiment

The principal component analysis and least squares method are compared to prove that the PCA has superiority. To simplify the calculation, the following model is built[11],as is shown in Figure 2.

a. Assume $D_1,...,D_5$ are mutually independent, and $D_i \sim N(0,0.4)$, d_i is the observed value of random variable D_i, $i = 1,...,5$;

b. x, y are variables, the observation value vectors of x, y are: $X = (x_1, x_2, x_3, x_4, x_5)'$,

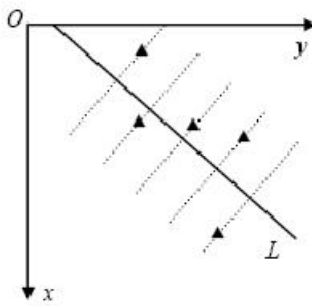


Figure 2 : Initial data generated randomly

$Y = (y_1, y_2, y_3, y_4, y_5)'$,

here,

$$\begin{cases} x_i = 5 + i - \dfrac{\sqrt{2}}{2} d_i \\ y_i = 6 + i + \dfrac{\sqrt{2}}{2} d_i \end{cases} \quad \textbf{(19)}$$

Obviously,the theoretical relationship between variables x and y is:

$$y = 1 + x + \varepsilon \quad \textbf{(20)}$$

Regression calculation is done based on the above data, using least squares regression method and Principal component analysis method separately, comparing the results with formula (13),then the regression error is got. Considering the slope of the regression line is large, small change in its position will cause a sharp change in slope, so it is not proper to use β_1 as the measurement variable of the error. In this paper the obliquity error of regression line $\Delta\alpha$ is used as the regression error metric. In addition, the intercepts β_0, β_1 of the regression lines are not independent, so discussing its error is nonsense. In conclusion, only one indicator $\Delta\alpha$ is selected to measure the performance of each regression method. (X, Y) is observed 30 times independently by using two methods for unary linear regression, and regression line inclinations are shown in TABLE 1,average error $\overline{\Delta\alpha}$ and mean absolute error $\overline{|\Delta\alpha|}$ are shown in TABLE 2,here,

$$\overline{\Delta\alpha} = \frac{1}{30}\sum_{i=1}^{30}(\alpha_i - 45^0) \quad \textbf{(21)}$$

$$\overline{|\Delta\alpha|} = \frac{1}{30}\sum_{i=1}^{30}|\alpha_i - 45^0| \quad \textbf{(22)}$$

TABLE 1 : Obliquity data of unary linear regressions(°)

LS	PCA	LS	PCA	LS	PCA
36.4614	36.7430	51.1583	52.5040	34.3672	36.5952
44.9712	46.0789	42.2432	43.4865	44.1101	45.4543
39.6035	47.3888	57.4840	57.6469	52.4824	57.6470
29.3945	30.1362	36.8264	38.4114	48.5504	50.4036
53.9081	54.6807	49.1087	55.8194	33.3938	35.8162
41.9061	44.3928	46.9750	49.2606	42.2317	44.1155
47.1428	50.2315	43.0093	44.6703	26.8489	27.5804
31.7569	33.8584	35.8210	36.1897	44.7187	47.1474
42.4422	43.0127	46.0340	52.6152	40.3706	41.4210
51.9599	52.1456	35.0894	36.0983	48.0692	51.4395

Annotation: LS means least squares method, PCA means principal component analysis.

TABLE 2 : Angle error data of unary linear regressions(°)

	LS	PCA
$\overline{\Delta\alpha}$	−2. 3854	−0. 2336
$\overline{\lvert\Delta\alpha\rvert}$	6. 2436	6. 5979

Seeing from TABLE 2, the average angle error $\overline{\Delta\alpha}$ in absolute value of principal component analysis is smaller than the least squares method, showing that principal component analysis has an advantage over least squares in terms of unbiased; the mean absolute error $\overline{|\Delta\alpha|}$ of principal component analysis is smaller than the method of least squares regression, indicating that principal component regression has an advantage over least squares method in terms of stability.

Application example

Annual reports of some Chinese coal mine listed enterprises are collected, among them, the data that non-permanent assets valued between 25 billion yuan and 150 billion yuan are shown in TABLE 3:

TABLE 3 : Annual reports data of some Chinese listed coal enterprises

Company(Co.,Ltd)	year	prime operating revenue, p (yuan)	non-permanent assets, *k* (yuan)
Henan Shenhuo Coal & Power	2012	27985000000	26417400000
Shanxi Lu'An Environmental Energy Development	2011	22426300000	27582110000
Shanxi Lu'An Environmental Energy Development	2012	13980400000	32268490000
Jizhong Energy Resources	2012	30072400000	26613100000
Yanzhou Coal Mining	2009	21500352215	45172821500
Yanzhou Coal Mining	2010	34844400000	54495300000
Yanzhou Coal Mining	2011	48768300000	76592900000
Yanzhou Coal Mining	2012	59673500000	96623500000
China Coal Energy	2006	28346700000	36712800000
China Coal Energy	2007	36823300000	41069800000
China Coal Energy	2008	52282566000	72945635000
China Coal Energy	2009	53729503000	83103856000
China Coal Energy	2010	71268400000	92494400000
China Coal Energy	2011	88872400000	129312600000
China Coal Energy	2012	87291700000	143319700000
China Shenhua Energy	2007	82107000000	121975000000
China Shenhua Energy	2008	107133000000	146466000000
China Shenhua Energy	2009	121312000000	164152000000

The scatter plots of the data in TABLE 3 are shown in Figure 3. All data are divided into two groups, scatter plots of values less than 100 billion yuan in non-permanent assets is represented by "*", used in the regression analysis;scatter plots of non-permanent assets values more than 100 billion yuan is represented by "o", and used in the validation of the regression results.

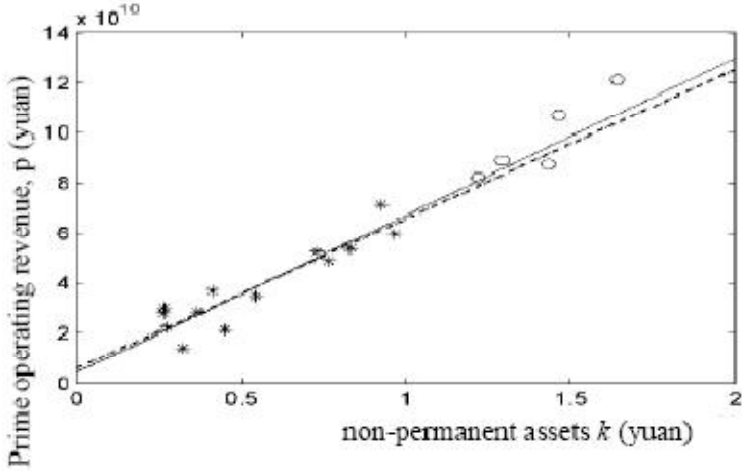


Figure 3 : Linear regression results

The result of the least squares regression based on the first set of data is:

$$\hat{p}_1 = 58.3537 + 0.5980k \quad \textbf{(23)}$$

The unit of above formula is one hundred million yuan. Least squares regression line is shown in dotted line in Figure 3.

The result of the principal component analysis regression based on the first set of data is:

$$\hat{p}_2 = 41.4890 + 0.6288k \quad \textbf{(24)}$$

The regression line of principal component analysis regression is the solid line in Figure 3.

For comparing the quality of two regression methods, the data that non-permanent value more than 100 billion yuan in TABLE 3 are used to validate regression results. Based on maximum likelihood estimation, the prime operating revenue of the second set of data are regarded as theoretical value, then calculate the error between regression error of predicted values and the theoretical values based on the two regression methods. Least squares regression error is denoted by ε_1, and the principal component analysis regression error is ε_2;the mean of regression error is denoted by $\overline{\varepsilon_2}$, the mean of the absolute value of regression errors is $\overline{|\varepsilon_2|}$. Results of regression errors are shown in TABLE 4.

TABLE 4 : Errors of regressions Unit: a hundred million

p	888.724	872.917	821.070	1071.330	1213.12
$\hat{p}_1$	831.662	915.426	787.782	934.241	1040.006
ε_1	-57.062	42.509	-33.288	-137.089	-173.114
$\hat{p}_2$	854.610	942.687	808.471	962.471	1073.681
ε_2	-34.114	69.770	-125.990	-108.859	-139.439

The mean of the regression errors by least square method is: $\overline{\varepsilon_1} = -71.609$.

The mean of the absolute value of regression errors by least square method is: $\overline{|\varepsilon_1|} = 88.612$.

The mean of the regression errors by principal component analysis method is: $\overline{\varepsilon_2} = -45.048$.

The mean of the absolute value of regression errors by principal component analysis method is: $\overline{|\varepsilon_2|} = 72.956$.

Mean of the regression errors can be regarded as systematic bias of regression method, as $\overline{\varepsilon_2} < \overline{\varepsilon_1}$,the systematic bias of principal component analysis method is smaller than least square method.

Mean of the absolute value of regression errors can be seen as standard deviation of regression errors, for $\overline{|\varepsilon_2|} < \overline{|\varepsilon_1|}$, the standard deviation of principal component analysis method is smaller than least square method, in another word, principal component analysis method is more stable than least square method.

CONCLUSION

Conventional principal component regression method is only used in the multiple linear regression, and only for the independent variable, the principal component obtained by analysis is regarded as independent variable, which are used in further regression analysis as well as causal variable. Unary linear regression method based on principal components analysis is described in this paper, which is extended to the area of unary linear regression, principal component analysis are carried out on both dependent variables and independent variables creatively. The regression line is attained based on the direction of the first principal component and data center. This method also has excellent property of the coordinate-independence.

Simulation experiment and application example both prove that the principal component analysis method used in linear regression is more precise and more stable than the least squares method. Principal component analysis method is an analytical solution which has the advantage of computation and has a broad application prospect.

The principal component analysis method and its application in a simple linear regression and its application is discussed in this paper, and the multiple regression will be discussed in another paper.

REFERENCES

[1] Liu Chun-Guo, Gao Song-Feng, Lu Xiao-Feng; Progress in the application of max autocorrelation factor method to remotely sensing imagery analysis[J]. Journal Of Henan Polytechnic University(Natural Science), **1**, 66-71 **(2012)**.

[2] Li Ming-Qi, Wu Xu; Partial ridge estimate of regression coefficients. Journal Of Henan Polytechnic University(Natural Science), **6**, 127-130 **(2011)**.

[3] Zhang Jian-Xiong, Zhao Guo-Qiang, Wu Xin-Hui; Point cloud coordinate transformation based on nonlinear least squares algorithm[J]. Journal Of Henan Polytechnic University(Natural Science), **4**, 62-66 **(2012)**.

[4] Cui Hong-Qing, Zhang Zhen-An, Li De-Jun; Prediction of deep gas content in coal mine based on simple nonlinear regression[J]. Journal Of Henan Polytechnic University(Natural Science), **6**, 6-9 **(2012)**.

[5] Xu Wei-Wei; Debt maturity structure of coal enterprises based on regression analysis. Friends of Accounting, **31**, 116-118 **(2011)**.

[6] Ning Yun-Cai; Risk Recognition and Management on Margin Trading in China. ICMSIS09 Proceedings, **09 (2009)**.

[7] Xu Wei-Wei; Risk Conversion of Debt Financing in the Coal Company. Zhengzhou:Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 5142-5145 **(2011)**.

[8] Xu Li-Li, Liu Shao-Wei; Establishing Prime System of Financial Management in Rural Enterprise. Procedings of 2007 International Conference on Agriculture Engineering, 11 **(2007)**.

[9] Sheng Zhou, Xie Shi-Qian, Pan Cheng-Yi; Probability Theory and Mathematical Statistics(Third Edition). Bei jing:Higher Education Press, 297 **(2001)**.

[10] Yu Xiu-Lin,Ren Xue-Song; Multivariate Statistical Analysis. Bei jing:China Statistics Press, 156-161, 239 **(1999)**.

[11] Xu Wei-Wei; A New Unary Linear Regression Method Based on Minimal Symmetrical Envelope Domain. Journal of Hangzhou Dianzi University, **34(2)**, 36-40 **(2014)**.