CSCanada
Progress in Applied Mathematics
Vol. 8, No. 1, 2014, pp. 29-33
DOI: 10.3968/5510

ISSN 1925-251X [Print]
ISSN 1925-2528 [Online]
www.cscanada.net
www.cscanada.org

A Simple and Practical Method of Calculating the Gini Coefficient

DONG Xiao[a],*; XU Feng[a]; ZHANG Shiqiang[a]

[a]Medical Information College, Chongqing Medical University, Chongqing, China.
* Corresponding author.

Address: Medical Information College, Chongqing Medical University, 400016, China. E-mail: math808@sohu.com
Received March 4, 2014/ accepted June 5, 2014/ Published online July 26, 2014

Abstract: The Gini coefficient is a way to describe socio-economic phenomena by mathematical model. Using an improved approximation regression method to estimate Gini coefficient in the model parameters. The regression accuracy of non-linear mathematical model seeked by improved method was significantly improved when compared with which seeked by the original approximation regression method. The normal equation derived from improved method which remains its convenient using advantages was just weighted from the normal equation derived from the original method.
Key words: Gini coefficient; Lorenz curve; Power function model

Dong, X., Xu, F., & Zhang, S. Q. (2014). A Simple and Practical Method of Calculating the Gini Coefficient. *Progress in Applied Mathematics, 8*(1), 29-33. Available from http://www.cscanada.net/index.php/pam/article/view/5510 DOI: http://dx.doi.org/10.3968/5510

1. INTRODUCTION

Early twentieth century Italy economist Gini, presented a judgment of equal distribution of index which called the Gini coefficient[1] $G \in [0,1]$. When $G=0$, it means absolute equality. The bigger G is the more unequal. When $G=1$, it means absolute inequality.

Market economy countries general standard to measure the income gap is: $G<0.2$ means that income is absolute equality; $G \in [0.2,0.3)$ means relative average; $G \in [0.3\text{-}0.4)$ means the income is reasonable; $G \in [0.4\text{-}0.5)$ means the income gap is a little big; $G \geq 0.5$ means the country has a wide income gap and it has polarized.

The Gini coefficient is a method with a mathematical model to describe the phenomenon of social economy. In general, the Lorenz curve is a nonlinear mathematical model. Because the nonlinear regression model is more complicate than the linear regression model, it is

not easy to calculate the regression parameters. To meet the needs of the actual situation in the premise, sometimes these approximate models[2] of the non-linear regression models are used to solve practical problems. To get the approximation of nonlinear regression model, we usually make variable substitution to nonlinear function, transform into a linear model, implement linear regression, and then reduce it to a nonlinear model. But this method will add the interference of information and loss of original information, sometimes affects the prediction accuracy of the nonlinear regression model[3]. We discuss the solution to solve these defects in the following sections

2. PRINCIPLE AND METHOD

2.1 Introduction to Nonlinear Mathematical Model

A nonlinear mathematical model can be expressed as the following form:

$$y=f(x_1, x_2,\ldots, x_m,; a_1, a_2,\ldots, a_h)+e\ , \tag{1}$$

in this fomula $e\sim N(0,s^2)$. The variable in this formula, $x=(x_1, x_2,\ldots, x_k,\ldots, x_m)\in R^m$, is a spot in m dimensional space and the parameter, $a=(a_1, a_2,\ldots,a_k,\ldots,a_h)\in R^h$, is a spot in h dimensional space and the dependent variable, $y\in R^1$, is a spot in one dimensional space. The multiple function of a is a nonlinear function of the variable x. In a nonlinear regression analysis, the first problem to be solved is how to get the best estimation in h dimensional space of a.

2.2 Approximate Estimation Method of Regression Model Parameters

In Statistics, there are some widely used nonlinear mathematical models like $y=Aa^{Bx}$, $y=Ax^B$, $y=A/(1+Ca^{Bx})$ and $y=[f(x)]^B\exp[Ag(x)]$.Assuming the data set of these mathematical model in m+1 dimensional space X-Y is $\{((x_1, x_2,\ldots, x_m)_i, y_i) \mid i=1, 2,\ldots, n\}$, the theoretical prediction data set correspongding to it is $\{((x_1, x_2, \cdots, x_m)_i, \hat{y}_i)\,|\, i = 1,2, \cdots, n\}$.

It's hard to find out the theoretical prediction data directly, so we usually make variable substitution to nonlinear function, transform into a linear model, implement linear regression, and then reduce it to a nonlinear model. We can use the variable substitution $z=F(y)$ to convert the data in m+1 dimensional space X-Y into the data in m+1 dimensional space X-Z, then we can find out the image collection of the data set in m+1 dimensional space X-Z as $\{[(x_1, x_2,\ldots, x_m)_i, z_i] \mid i=1,2,\ldots, n\}=\{[(x_1, x_2,\ldots, x_m)_i, F(y_i)] \mid i=1,2,\ldots, n\}$. Then its theoretical prediction data set in m+1 dimensional space X-Z is $\{((x_1,x_2,\cdots,x_m)_i,\hat{z}_i)\,|\, i=1,2,\cdots,n\}=\{((x_1,x_2,\cdots,x_m)_i,F(\hat{y}_i))\,|\, i=1,2,\cdots,n\}$.

According to the experiment data set to find out the corresponding nonlinear mathematical model, many literatures usually get the residual sum of squares in m+1 dimensional space X-Z

$$S_1=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2\ . \tag{2}$$

Then find out the best estimation $\hat{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_l)\in R^l$ of parameter a in h dimensional space by using the least square method. At last put the best estimation into Equation (1), we can get the nonlinear mathematical model we need.

But adopting this method to get the nonlinear mathematical model has hidden serious

defect which is the residual sum of squares in m+1 dimensional space X-Z and the parameter a we get in h dimensional space when S_1 is minimum are uncertain to make sure that the residual sum of squares in m+1 dimensional space X-Y is minimum:

$$S_2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{3}$$

This is exactly the defect which causes the error in the regression parameter of the nonlinear mathematical model; sometimes it can make the model completely inefficient.

2.3 Improvement of the Parameter Estimation Method

From above analysis we can see that if we want to lead out a good nonlinear mathematical model, we should use data mining method and make full use of the information from initial data to make sure the residual sum of squares in m+1 dimensional space X-Y is minimum. Expand the function which contains unknown parameter a in h dimensional space on the y_i by Taylor series, then we can find out:

$$\hat{z}_i = F(\hat{y}_i) = z_i + F'(y_i)(\hat{y}_i - y_i) + \frac{1}{2}F''(y_i)(\hat{y}_i - y_i)^2 + \cdots$$

When $\hat{y}_i \to y_i$, Omit the infinitesimal all higher order infinite small can get:

$$(\hat{y}_i - y_i) \approx \frac{(\hat{z}_i - z_i)}{F'(y_i)}. \tag{4}$$

So we can get the approximate expression of the residual sum of squares in m+1 dimensional space X-Y.

$$S_2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \approx \sum_{i=1}^{n} \frac{(z_i - \hat{z}_i)^2}{[F'(y_i)]^2}. \tag{5}$$

Use the least square method to (5), find out the normal equations corresponding to the nonlinear mathematical model and find out the beat estimation $\hat{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_l) \in R^l$ of the parameter a in h dimensional space, at last put the best estimation into (1) to get the nonlinear mathematical model in the original variables dimensional space X-Y.

3. APPLICATION OF PARAMETER ESTIMATION METHOD

Use simple power function to simulate the Lorenz curve by this method, then the nonlinear mathematical model of Lorenz curve based on power function is $y=x^{\alpha}$. Assume variable substitution $z=F(y)=\ln y$, use variable substitution to convert the data in 2 dimensional space X-Y into the data in 2 dimensional space X-Z, then find out the parameter a. So we can find out that the image set in 2 dimensional space X-Z of the experimental data set which comes from the nonlinear mathematical model in 2 dimensional space X-Y which has not been fitted is: $\{(x_i, z_i) \mid i=1, 2, \ldots, n\} = \{[x_i, F(y_i)] \mid i=1, 2, \ldots, n\}$. Then the image set of the predicted value of the theoretical data corresponding to the parameter a is $\{(x_i, \hat{z}_i) | i = 1,2,\cdots,n\} = \{(x_i, F(\hat{y}_i)) | i = 1,2,\cdots,n\}$. Through Equation (5) and $F'(y)=[\ln y]'=1/y$, we can find out: $S_2 \approx \sum_{i=1}^{n} y_i^2 (\ln y_i - \alpha \ln x_i)^2$. Derivate the above equation, we can get: $\frac{dS_2}{d\alpha} \approx -2\sum_{i=1}^{n} [y_i^2 \ln y_i \ln x_i - \alpha y_i^2 (\ln x_i)^2]$. Assume the above formula is equal to zero,

then

$$\hat{\alpha} = (\sum_{i=1}^{n} y_i^{\,2} \ln y_i \ln x_i) \Big/ (\sum_{i=1}^{n} y_i^{\,2} (\ln x_i)^2) .$$

4. THE EMPIRICAL ANALYSIS

According to the survey data of the 56,094 households living conditions of urban residents in China Statistical Yearbook 2007, there are seven groups divided In accordance with the revenue form low to high order. They are the lowest income families, low income, under the average income, middle income, above the average income, high income households, and the highest income households. Table 1 is cumulative percentage of the population data and cumulative percentage of the revenue data.

Table 1
Cumulative Percentage of the Population *x* and the Revenue *y*

x	11.22	22.09	43.16	63.04	82.01	91.16	100
y	3.38	8.42	21.71	38.81	61.27	76.01	100

The Gini coefficient calculated by geometry estimation method put forward by *Nuria*[5] in Reference [4] is G=0.3952; The Gini coefficient calculated by Beta model put forward by *Kakwani*[6] is G=0.3399; The Gini coefficient calculated by Pareto distribution function put forward by Chen Xiru [1] and Chen Zhiqi[2] is G=0.3449.

Use the method put forward in this article to calculate the Gini coefficient, we can get $\hat{\alpha}$ $\hat{\alpha}$=1.990481, the equation of Lorenz curve is $y=x^{1.990481}$ and G=0.3312. So we can see that the Gini coefficient calculated by this method is basically agreeing with the Gini coefficient calculated by the three methods we talked above.

5. DISCUSSION

This paper puts forward a method for parameter estimation of Gini coefficient model and uses the improved approximate regression method to estimate parameter*a*. Both the improved method and the original one have converted the data in *m*+1 dimensional space *X-Y* into the data in *m*+1 dimensional space *X-Z*. But the original method used Equation (2) the residual sum of squares in *m+1* dimensional space *X-Z* to get the nonlinear mathematical model. The improved method used Equation (3) the residual sum of squares in *m*+1 dimensional space *X-Y* to get the nonlinear mathematical model. Then use the best estimate of the value of the least squares method to calculate the parameters $\hat{\alpha}$ of *h*-dimensional space, and substituted into Equation (1) to obtain the nonlinear mathematical model requested. It is this difference that makes the improved method more accurate than the original one.

Use the power function model to simulate the Lorenz curve ,from which we get the Gini coefficient G that agrees with that calculated by software Eviews5.5 with Beta model put forward by Kakwani (1986), which calculated by software Eviews5.5 with Pareto distribution function put forward by Chen (2004) and Chen (2006). From above we can see that using power function to simulate the Lorenz curve is feasible. In addition, the Beta model contains three parameters and the Pareto distribution function given by Chen (2004) and Chen (2006) contains two parameters, of which the workload is so relatively large that the giniral need Eviews5.5 software to complete. Using a simple power function to

simulate Lorenz curve model has only one parameter. It just needs a calculator with simple logarithmic function to complete the work, very convenient for using.

REFERENCES

[1] Chen, X. R. (2004). Gini coefficient and its estimation. *Statistical Research,* (8), 58-60.

[2] Chen, Z. Q., & Chen, J. D. (2006). Gini coefficient and its estimation. *Journal of Beijing University, 42*(5), 613-618.

[3] Zhang, S. Q. (2002). Approach on the fitting optimization index of curve regression. *Chinese Journal of Health Statistics, 19*(1), 9-11.

[4] Wang ,Y. M. (2010). Comparison of the calculation method of Gini coefficient. *Statistics and Decision,* (5), 157-159.

[5] Nuria, B. P. (2003). *Approximation of Gini index from grouped data*. Working Paper.

[6] Kakwani, T. (2007). An analysis of the impancts of development on Gini inequality using grouped and individual observations: Examples from the 1998 vietnamese household expenditure data. *Journal of Asian Economics, 18*.