OPEN ACCESS

EURASIA Journal of Mathematics Science and Technology Education
ISSN: 1305-8223 (online) 1305-8215 (print)
2017 13(8):5765-5773
DOI: 10.12973/eurasia.2017.01026a

# A Newer Equal Part Linear Regression Model: A Case Study of the Influence of Educational Input on Gross National Income

Wen-Tsao Pan
School of Business, Guangdong University of Foreign Studies, China

Received 10 February 2017 ▪ Revised 4 June 2017 ▪ Accepted 1 August 2017

**ABSTRACT**
Linear Regression Model (LRM) is not only a time – hornored reserch method but a simple and essential analytical technique for social science researchers. However, it cannot reveal the meaning of the extrme value included in genuine data, which constitutes the major concern for researchers in social scineces. To solve this problem, most researhcers turned to quantile regression model. This model, hoever, is not only obscure because it divides data by quantile but difficult to perform because it needs special software. The study therefore proposed a new concept as well as method: equal part linear regression model (EPLRM) which divides the sample data into equal parts and builds LRM on each part so that the research can both observe the data distribution of sample data within each part and compare the results with that of LRM. As the case study shows, the poverty of fiscial expenditure on education would decrease Gross Nation Income (GNI) greatly and the promotion of educational input of private schools and social donation would boost the increase of GNI to a lage degree.

**Keywords:** extreme value (EV), linear regression (LR), equal part linear regression (EPLR), gross national income (GNI)

## INTRODUCTION

Linear Regression Model (LRM) constitutes an important method for researchers in social sciences to conduct data analysis. Literatures so far are mainly focused on education and learning (He Xianying & Wang Li, 2014; Liu Kehui et al., 2014; Li Yongci, 2008; Luo Min et al., 2009). In statistics, LRM is an approach for modelling the relationship between one or more independent variables (X) and dependent variables (Y) by using the least squares in linear regression equation. It is a practical technique to predict the financial market. Numerous extensions of LRM have been developed. Applied to data in normal distribution, LRM is able to produce ideal analytical and predicting results. However, besides normal data, there exist some extreme values in genuine data set, which act as the major concern for social science researchers. If these extreme values are dealt with LRM and simplified as mean value, the research results may become unreliable.

This problem is more often than not tacked by Quantile Regression put forward by Koenker in 1978. Some researchers (Bai Xuemei & Li Ying, 2014; Shang Jun et al., 2013; Huang Fenfen, 2015; Pan Wen Tsao et al., 2016) have studied in this respect. Quantile, however, is by no means a household concept. Quantile Regression itslef is similarly complex and obscure for the public. Moreover, the modeling of Quantile Regression is somewhat demanding since it requires of special software. Therefore, the author proposes a new model -- the Equal Part Linear Regression Model (EPLRM). The new approach first divides the data set into equal parts and then conducts the


**Correspondence:** Wen-Tsao Pan, *Professor, School of Business, Guangdong University of Foreign Studies, China. Address to No.2 North Baiyuan Avenue, Baiyun District Guangzhou, China. Tel: +286-20-36207878.*
teacherp0162@yahoo.com.tw

**State of the literature**

- Multiple regression analysis is an important analyzing method, and it is now widely applied in various fields (He Xianying & Wang Li, 2014; Liu Kehui et al., 2014). In the statistics, multiple linear regression model (LRM) uses the least squares function of the linear regression equation to model the relation between one or several independent variable (X) and dependent variable. It is a feasible finance market predicting method with high practical value.
- Koenker and Bassett (1978) proposed the quantile regression method. Different from the conventional linear models that predict the mean of the explained variable given a specific value of each explanatory variable, a quantile regression model predicts the value of the explained variable at a specific quantile of the explained variable giving a specific value of each explanatory variable.

**Contribution of this paper to the literature**

- In this paper, a new regression modeling method is proposed. In this method, data division is used to observe and analyze the detailed data distribution, so as to improve the defects of standard linear regression.
- Compared to the quantile regression model which is widely used now, this method is more comprehensible and analysis will be easier.
- The study proposed a new technique to build LRM which divides sample points into equal parts and moves the equal part from left to right to build EPLR lines and observe and analyze the data distribution of smaller parts to improve the effect of stand linear regression model.

linear regression modeling on this so that researchers can observe the trend of each equal part and compare it with the results of general LRM. We consulted related literature (Zhang Yi, 2007; Chen Jianbao & Dai Ping Sheng, 2007), selected real cases, collected the data of China's educational input and Gross National Income (GNI) over the years and analyzed with EPLRM.

The paper first introduces the research motivation, research purpose and related literature then illustrates the EPLRM and its testing and conducts an empirical study based on EPLRM and genuine data and finally concludes with research results and suggestions.

## RESEARCH METHOD

### Equal Part Linear Regression Model

If $y$ is a continuous independent variable, then the LRM after modeling can be stated as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where the $\varepsilon_i = \mu = 0$, the standard deviation $\sigma^2$ will have normal distribution. Since linear regression is the straight line to find the least error sum of squares by using least squares, it can be expressed as:

$$\min \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2$$

The solution of $\beta_0$ and $\beta_1$ is called ordinary least squares estimator (OLS estimator) which is expressed by $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \bar{\beta}_0 \bar{x}$$

and the coefficient of determination, denoted $R^2$, is the most common fitting indicator to measure the proportion of the variance in the dependent variables ($y$) that is predictable from the independent variable(s) ($x$). The formula is

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\bar{y}}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where the higher the $R^2$ is, the higher the ratio of the explained variance to the total variance is and the better the model fits the data. Thus when $R^2 = 1$, regression sum of squares equals total sum of squares and the regression model can explain all variability in $y$, which is denoted perfect fit. When $R^2 = 0$, the residual sum of squares equals total sum of squares where the fitted model fails to explain the variability in $y$. Suppose the confidence level of $(1 - \alpha)$, the confidence interval of $\beta_i$ will be

$$\left(\hat{\beta}_i - t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_t}, \hat{\beta}_i + t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_t}\right), i = 0,1$$

EPLRM instead divides linear data into "$\tau$" equal parts and models respectively. **Figure 2** divides nine data points into three equal parts, each part including other three data points. Since the trend of these three data points vary from part to part, if analyzed by standard LRM, the results may become unreliable. Therefore, we can fit LRMs respectively to these three data points of each part. These three EPLR formulas can be expressed as

$$y_i^{\tau} = \beta_0^{\tau} + \beta_1^{\tau} x_i^{\tau} + \varepsilon_i^{\tau}$$

Least squares estimation

$$\hat{\beta}_0{}^{\tau} = \frac{\sum_{i=1}^{n}(x_i{}^{\tau} - \bar{x}^{\tau})(y_i{}^{\tau} - \bar{y}^{\tau})}{\sum_{i=1}^{n}(x_i{}^{\tau} - \bar{x}^{\tau})^2}$$

$$\hat{\beta}_1{}^{\tau} = \bar{y}^{\tau} - \bar{\beta}_0{}^{\tau}\bar{x}^{\tau}$$

the coefficient of determination as well as confidence interval

$$\left(\hat{\beta}_i{}^{\tau} - t_{\alpha/2} \times s_{\hat{\beta}_t}\hat{\beta}_i{}^{\tau} + t_{\alpha/2} \times s_{\hat{\beta}_t}\right), i = 0,1$$

where "$\tau$" can be inserted in particular spaces.

## Distance and Sample Point EPLRM

In accordance with different research questions, EPLR can be put into two categories: Distance Equal Part Linear Regression (DEPLR) and Sample Point Equal Part Linear Regression (SPEPLR). **Figure 1** is an example of DEPLR where the distance between starting point to cut point 1, cut point 1 to cut point 2 and that after point 2 is equal. From starting point to pint 1, there exists 4 sample points, point 1 to point 2, 5 sample points, after point 2, 3 sample points. Keep in mind that when use DEPLR, the sample points in each equal part must equals or exceeds 3,

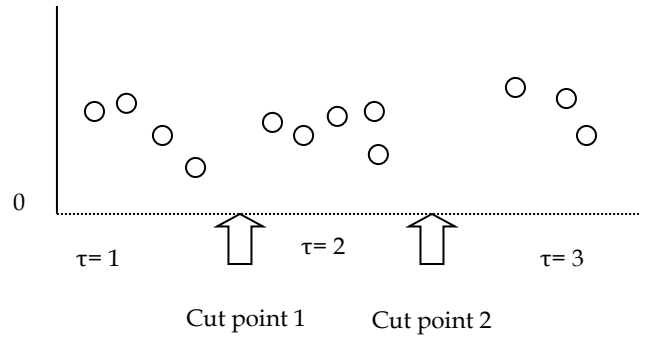


**Figure 1.** Three EPLRMs (Distance Equal Part)

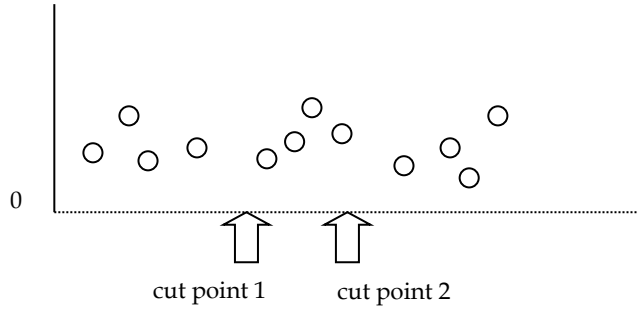


**Figure 2.** Three EPLRMs (Sample Points)

**Table 1.** Test Data of EPLRM

| **X** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Y** | 3 | 6 | 9 | 10 | 11 | 12 | 13 | 14 | 16 | 18 | 21 | 25 |

so that the LRM can be established. DEPLR can be used when sample points cannot be dived equally (for instance, the multiple of 2, 3 and so on)

**Figure 2** is the example of SMEPLR where the distance of each part is distinctive but the number of sample points within each part is equal (namely 4 sample points). Generally speaking, SPEPLR is comparatively easier to be realized in that it only divides sample points into equal parts. DEPLR instead first divides the data set into several parts with equal distance and then calculate to ensure the sample points within each part equal or exceed three. Therefore, DEPLR is more often than not used in special cases.

## Equal Part Linear Regression Model Testing

Given a data set of 12 sample points and ordered pair of X, the test data are as follows (see **Table 1**).

LRM and three EPLRMs are simultaneously built and the regression formulas of them are as follows:

$Y = 2.257 + 1.678X \qquad R^2 = 0.955$

if $\tau = 1; Y = 1 + 2.4X \qquad R^2 = 0.980$

if $\tau = 2; Y = 6 + X \qquad R^2 = 1$

if $\tau = 3; Y = -11.5 + 3X \qquad R^2 = 0.978$

As shown by the three EPLR formulas, the direction and tilt degree of the fitted regression line is different from that of the LR and the explanatory power ($R^2$) of the former is stronger than that of the latter. In **Figure 3**, the horizontal axis is X, vertical axis Y, solid line the LR line and three dashed lines EPLR lines. It is found again that the direction and tilt degree of the three fitted regression lines are all distinct from that of the LR lines. **Figure 3** also shows that, when X is less than 2, the result of the LR was overestimated than that of the first EPLR formula ($\tau = 1$) and when X equals or exceeds 2, the results was underestimated. When 4 < X<6, the result of was underestimated than that of the second EPLR formula ($\tau = 2$) and when X equals and exceeds 6, the result was overestimated. When 8 < X <10, the result was similarly overestimated than that of the third EPLR formula ($\tau = 3$) and when X equals and exceeds 10, the result was underestimated.

Overall, it is better to use EPLR than LR to predict data distribution.

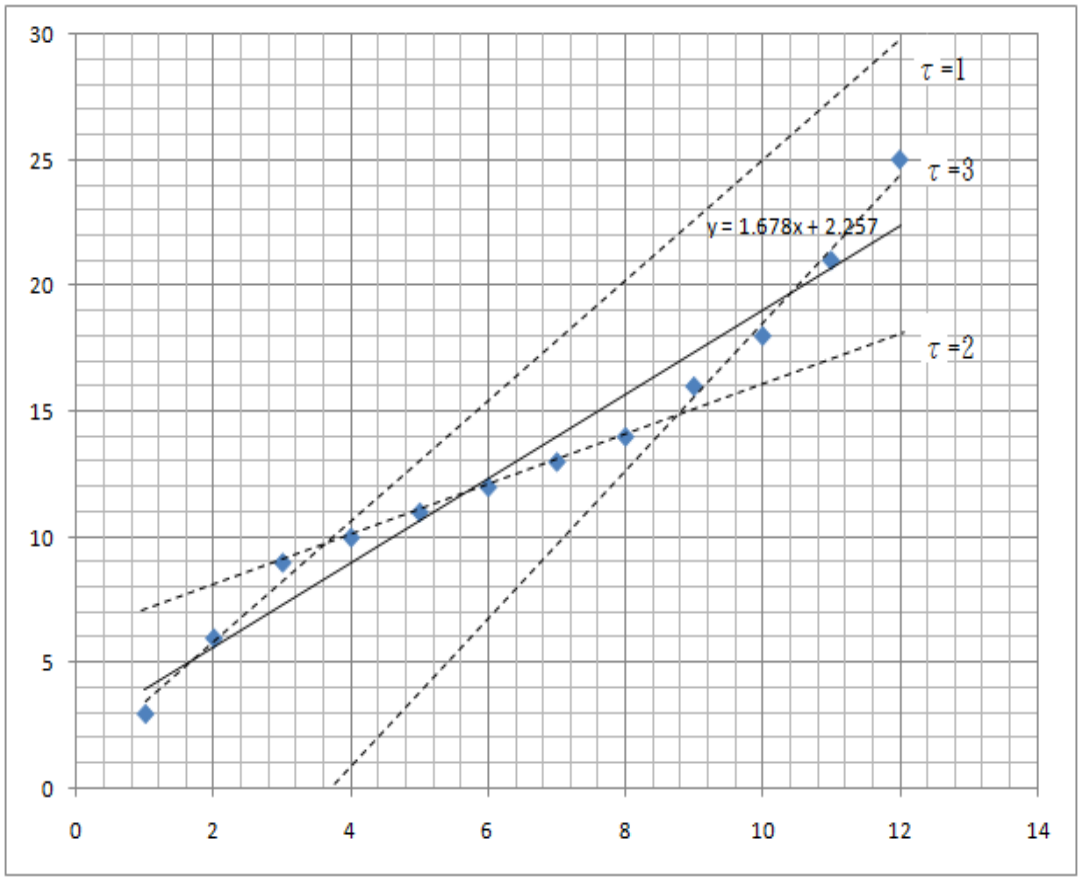


**Figure 3.** LR Line and Three EPLR Lines

**Table 2.** Sample Data of EPLRM

| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| X1 | 2562.61 | 3057.01 | 3491.40 | 3850.62 | 4465.86 | 5161.08 | 6348.36 | 8280.21 |
| X2 | 85.85 | 128.09 | 172.55 | 259.01 | 347.85 | 452.22 | 549.06 | 80.93 |
| X3 | 113.96 | 112.89 | 127.28 | 104.59 | 93.42 | 93.16 | 89.91 | 93.06 |
| Y | 99066.1 | 109276.2 | 120480.4 | 136576.3 | 161415.4 | 185998.9 | 219028.5 | 270844 |
| **No.** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | |
| X1 | 10449.63 | 12231.09 | 14670.07 | 18586.70 | 23147.57 | 24488.22 | 26420.58 | |
| X2 | 69.85 | 74.98 | 105.43 | 111.93 | 128.18 | 147.41 | 131.35 | |
| X3 | 102.67 | 125.50 | 107.88 | 111.87 | 95.69 | 85.54 | 79.67 | |
| Y | 321500.5 | 348498.5 | 411265.2 | 484753.2 | 539116.5 | 590422.4 | 644791.1 | |

## CASE STUDY

### Three Equal Part Linear Regression Model Building

This study focused on the GNI of the Chinese mainland, treated the fiscal expenditure on education (X1), the educational input of private schools (X2), social donation (X3) and GNI (Y) as the objects of study and explored the influence of governmental and private educational input on the GNI. The following sample data from 2000 to 2014 are extracted from China Statistical Yearbook (See **Table 2**).

**Table 3.** Statistical result of the four LRMs

| | LRM $R^2$=0.989 | | | EPLRM $\tau$=1 $R^2$=0.996 | | | EPLRM $\tau$=2 $R^2$=0.999 | | | EPLRM $\tau$=3 $R^2$=0.977 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Stat.** | **Conf.** | **T** | **Sig.** | **Conf.** | **T** | **Sig.** | **Conf.** | **T** | **Sig.** | **Conf.** | **T** | **Sig.** |
| X1 | 21.834 | 21.826 | *** | 9.573 | 0.412 | - | 26.562 | 49.025 | ** | 17.222 | 2.269 | - |
| X2 | -2.506 | -0.045 | - | 159.964 | 0.859 | - | -7.350 | -1.782 | - | -730.759 | -0.424 | - |
| X3 | -213.835 | -0.355 | - | -78.144 | -0.224 | - | -851.356 | -11.664 | * | -1411.86 | -0.582 | - |

Note: *indicates 10% significance level; **indicates 5% significance level; ***indicates 1% significance level.

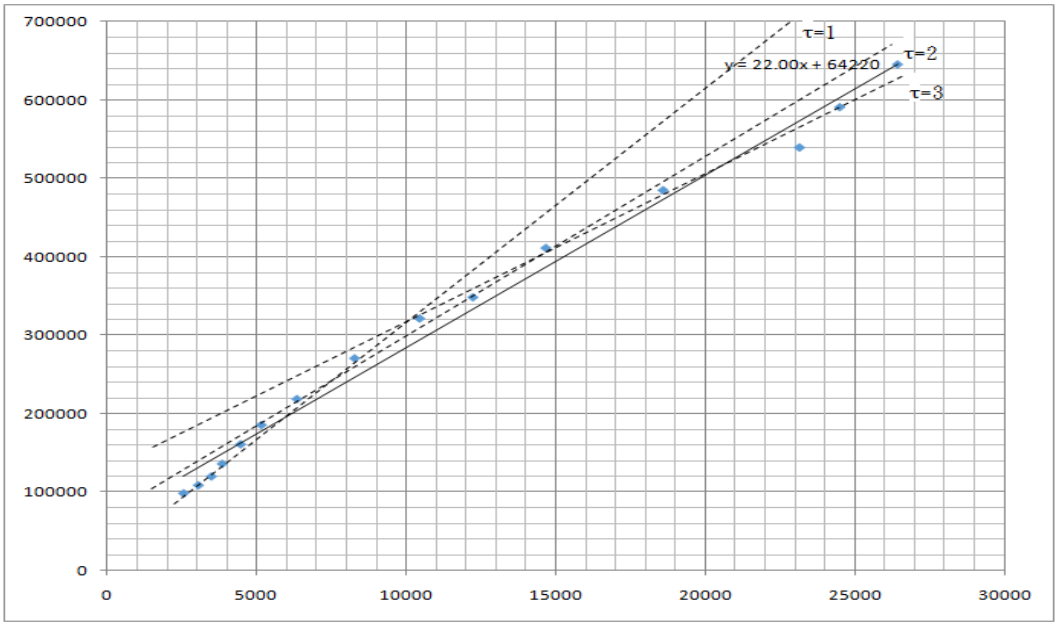


**Figure 4.** The LR Line and EPLR Lines between X1 and Y

The study first analyzed the data set with SPEPLR and then explored the influence of X1, X2 and X3 respectively on GNI (Y). Practically, the study sorted the data of X1 by ascending counts and meantime built LRM and three EPLRMs. Since there are 15 sample points, each equal part is divided with five sample points. **Table 3** is the analytical results of LRM and three EPLRMs based on X1, X2 and X3. It is found that facial expenditure on education (X1) in LRM has had an obvious effect on Y. In EPLRM, when $\tau = 2$, the influence is significant, the X1 in the middle equal part has had an obvious effect on the Y and this was not applicable to the situation when $\tau = 1$ and $\tau = 3$. There exists no significant influence between the stock of fixed asset (X1) in the lower and higher equal parts and the GNI (Y), which cannot be observed through LRM. **Table 4** is the result of the LRM and three EPLRMs between X1 and Y, where the result may be overestimated when dealt with EPLR in the first equal part; the result may be underestimated in the second equal part and overestimated in the third equal part.

As **Table 3** shows, in the LRM no obvious influence of social donation (X3) can be observed on GNI (Y), but in EPLRM, when $\tau = 2$, the influence is significant, namely, X3 has had an obvious effect on Y and when $\tau = 1$ or $\tau = 3$, the influence is not significant.

**Table 4** is the result of difference test between groups where F test is run on Y/X1 and Y/X2 of each equal part. It is found that in EPLRM there exists significant difference between the first part and second part of X1, the first part and the second part, the second part and the third part of X2 as well as the second part and the third part, the first part and the third part of X3. As shown in **Table 5**, the difference test of the coefficient of variation, there

**Table 4.** F Test of the Sample Data between Equal Parts

| Variable (s) | $\tau_1 - \tau_2$ | | $\tau_2 - \tau_3$ | | $\tau_1 - \tau_3$ | |
|---|---|---|---|---|---|---|
| | F_value | Sign. | F_value | Sign | F_value | Sign |
| X1 | 0.0729 | ** | 4.8550 | - | 0.3537 | - |
| X2 | 0.0078 | *** | 26.3906 | ** | 0.2045 | - |
| X3 | 1.5823 | - | 0.0323 | *** | 0.0511 | ** |

Note: *indicates 10% significance level; **indicates 5% significance level; ***indicates 1% significance level.

**Table 5.** F Test of the Coefficient between Equal Parts

| Variable (s) | $\tau_1 - \tau_2$ | | $\tau_2 - \tau_3$ | | $\tau_1 - \tau_3$ | |
|---|---|---|---|---|---|---|
| | F_value | Sign. | F_value | Sign | F_value | Sign |
| X1 | 35.1412 | ** | 0.3441 | - | 18.1590 | * |
| X2 | 0.1280 | - | 0.1950 | - | 0.0411 | ** |
| X3 | 24.7515 | ** | 0.6681 | - | 19.5858 | ** |

Note: *indicates 10% significance level; **indicates 5% significance level; ***indicates 1% significance level.

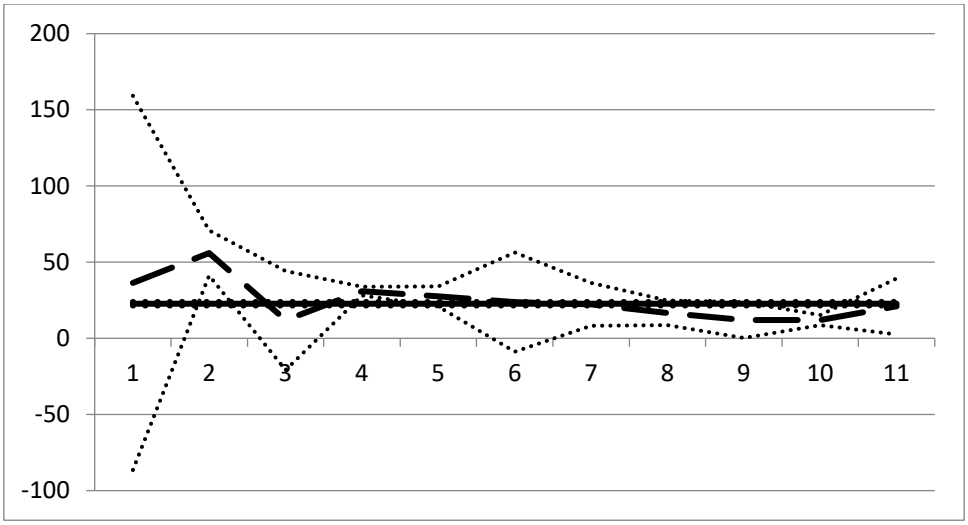


**Figure 5.** LR and EPLR between X1 and Y

exists significant difference between the first part and the second part, the first part and the third part of the coefficient of X1, between the first part and the third part of the coefficient of X2 and between the first part and the second part, the first part and the third part of X3. Detailed analysis will be shown in the following part.

## Creation of Equal Part Linear Regression Lines and Confidence Interval

If we build one regression model on the basis of all the five sample points in one equal part to estimate the trend of the 15 sample points, a serious error will occur. The author therefore moves the five sample points to the right by one X coordinate, builds a regression line each time and creates the coefficient as well as confidence interval of X1, X2 and X3 (as shown in **Figure 5**, **6** and **7**) to avoid potential errors.

**Figure 5** is the trend of the coefficient of X1 where the heavy line in the middle is LRM, the horizontal thick dash line placed up and down the confidence interval of LRM, the irregular thick dash line in the middle EPLRM and irregular fine dash line up placed up and down the confidence interval of EPLRM. From **Table 3** we know that when X1 is located in the second equal part ($\tau = 2$), there exists significant difference. From **Table 4** we know that there exists significant difference between the first part ($\tau = 2$) and the second part ($\tau = 2$). As shown in **Figure 5**, the results would be underestimated when dealt with LRM in the first part and the second part and overestimated

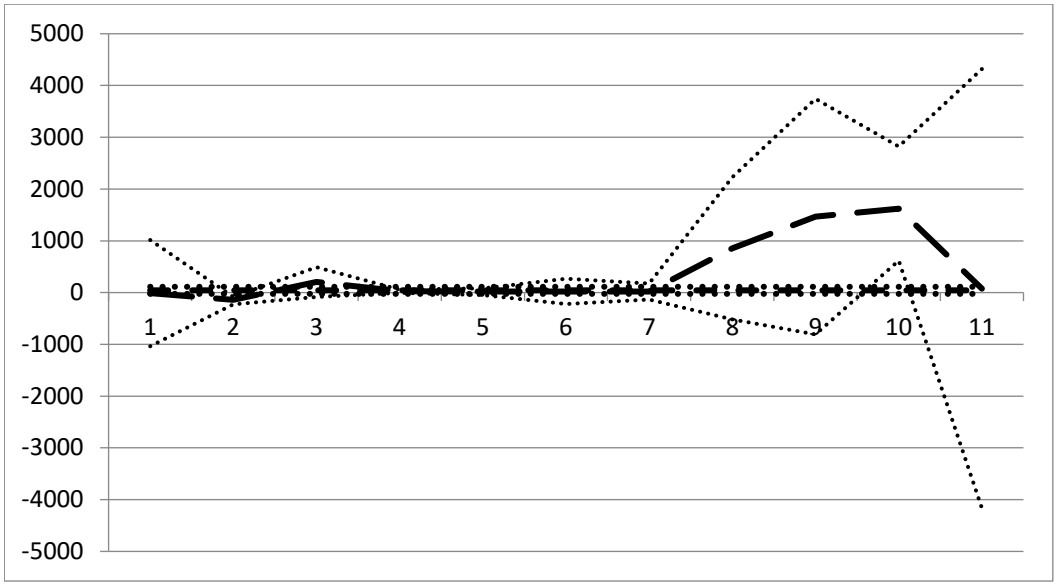


**Figure 6.** LR Line and ELR Lines between X2 and Y

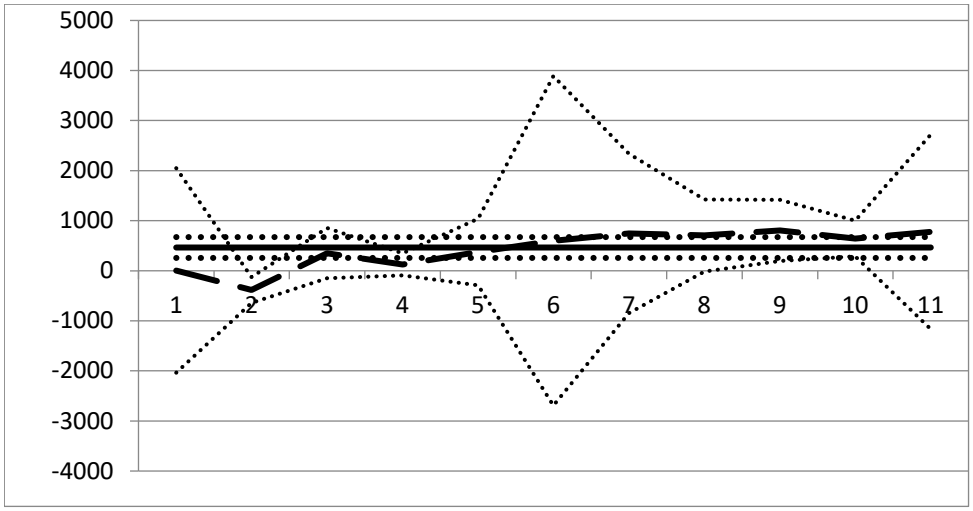


**Figure 7.** LR Line and EPLR Lines between X3 and Y

in the third part, which indicates that the poverty of fiscal expenditure will significantly lower the GNI. Government in this case is advised to increase the fiscal expenditure on education to boost GNI.

From **Table 4** and **5**, we know that there exists significant difference between the first part (τ=1) and the second part ($\tau = 2$) of X2 and from **Table 4** we also know there exists significant difference between the second part ($\tau = 2$) and the third part ($\tau = 3$). Besides, as shown in **Figure 6**, the results will be overestimated when dealt with LR in the first part of X2 and underestimated in the third part, which indicates that the investment in the educational input of private schools (X2) will greatly promote the increase of GNI.

Similarly, when observing X3 in **Table 4** and **5**, we may find significant difference exists between each equal part. Besides, as shown in **Figure 7**, the results would be overestimated when analyzed by LR in the LRM first equal part and underestimated in the third equal part, which indicates that although the poverty of social donation (X3) won't decrease GNI, but the investment in social donation will greatly promote the increase of GNI.

## CONCLUSION AND SUGGESTIONS

The study proposed a new technique to build LRM which divides sample points into equal parts and moves the equal part from left to right to build EPLR lines and observe and analyze the data distribution of smaller parts to improve the effect of stand linear regression model. Compared with the widely-used quantile regression model, EPLRM is easier to get across and to perform. Since the sample data of this case study is intended for simple test, the author only chose the data of China Statistical Yearbook from 2000 to 2014. However, in case of building EPLRM again, it is advised to divide the equal data by such variables as X or Y and to adjust the number of the sample points within each equal part to meet the research purpose. Last but not least, since it is rather consuming and demanding when, the building of EPLRM is hoped to be encoded and realized by software because it will become rather consuming and demanding when the sample data is huge enough.

## ACKNOWLEDGMENTS

The research is supported by the National Natural Science Foundation of China (71673064), Humanities and Social Sciences Planning Project of the Ministry of Education (13YJC630240), Soft Science Project of the Science and Technology Pro-gram of Guangdong Province (2013B070206058, 2015A070704054), Outstanding Youth Fund Project of Education Department of Guangdong Province (2014WTSCX040), Soft Science Project of the Science and Technology Program of Guangzhou (2014Y430009).

## REFERENCES

Chen Jianbao & Dai Pingsheng (2007). Empirical Analysis of Spatial Characteristics between Education and GDP in Different Regions of China. *Education & Economy*, (3), 20-25.

Equal Part Linear Regression programming in R. (n.d.). Retrieved from http://eplrm.byethost31.com/

Harding, M., & Lamarche, C. (2009). A quantile regression approach for estimating panel data models using instrumental variables, *Economics Letters, 104*(3), 133–135. doi:10.1016/j.econlet.2009.04.025

Koenker, R. W. & Bassett, G. (1978). Regression Quantile, *Econometrica, 46*(1), 33-50. doi:10.2307/1913643

Liu Kehui, Yan Xiaojun & Zhao Yishu et al. (2014). Experimental Study of Human Depth Perception Based on SPSS. *Journal of North China Institute of Science and Technology, 11*(8), 62-66.

Meligkotsidou, L., Vrontos, I. D., & Vrontos, S. D. (2009). Quantile regression analysis of hedge fund strategies, *Journal of Empirical Finance, 16*(2), 264–279. doi:10.1016/j.jempfin.2008.10.002

Pan, W. T., Huang, C. E., & Chiu, C. L. (2016). Study on the performance evaluation of online teaching using the quantile regression analysis and artificial neural network. *Supercomputing*, (72), 789–803. doi:10.1007/s11227-015-1599-1

Pan, W. T., Leu, Y. H., Zhu, W. Z., & Lin, W. Y. (2017). A Data Mining Approach to the Analysis of a Catering Lean Service Project, *Intelligent Automation & Soft Computing, 23*(2), 243-250. doi:10.1080/10798587.2016.1203564

Watanabe, S., Nakamura, A., & Juang, B. H. (2014). Structural Bayesian Linear Regression for Hidden Markov Models, *Journal of Signal Processing Systems, 74*(3), 341-358. doi:10.1007/s11265-013-0785-8

Yu. R. Y. and Rue, H. (2011). Bayesian inference for additive mixed quantile regression models, *Computational Statistics & Data Analysis, 55*(1), 84-96. doi:10.1016/j.csda.2010.05.006

Zhang Yi. (2007). An Empirical Analysis on Educational Input and GDP Development Cointegration. *Journal of Changchun Normal University, 26*(8), 20-23.