We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter 5

Fitting Truncated Mode Regression Model by Simulated Annealing

Maoxi Tian, Jian He and Keming Yu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/66070

Abstract

Like mean, median, and standard deviation, mode as the value that appears most often in a set of data is an important feature of a distribution. The numerical value of the mode is the same as that of the mean and median in a symmetric distribution but may be very different in a highly skewed distribution. Mode regression, which models the relationship between the mode of a dependent variable and some covariates, was first introduced by Lee in terms of truncated dependent variables. Some modifications of the truncated mode regression have been proposed recently. However, little progress is made on the computation or algorithm of fitting a mode regression due to an NP-hard optimization problem. In this paper we first introduce the popular simulated annealing (SA) to solve the truncated mode regression optimization. Experiments with simulations compare favorably to SA. Then, a mode regression with the proposed algorithm is applied to explore the typical income structure of China. We also compare the income returns to gender, education, experience, job sector, and district between the majority of workers with typical income and the workers with mean, middle income via comparison between mode regression, mean regression, and median regression.

Keywords: income inequality, Lee's estimate, median regression, mode, mode regression fitting, simulated annealing algorithm, truncated variable

1. Introduction

Mode, the most likely value of a distribution, has wide applications in biology, astronomy, economics, and finance. For example, in the archeology, many practical questions often focus on "Which era the biological species most likely to survive in according to the biological fossils?" In such cases, mode regression provides a convenient summary of how the repressors

INTECH
open science | open minds

© 2017 The Author(s). Licensee InTech. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

affect the conditional mode. Income inequality is of growing concern to people in rich as well as poor countries. This inequality may result in problem in social stability in a region or a country. But using average income as a statistical measurer is largely affected by a small number of richest who earned a high proportion of all income. Mode is the right statistic to measure the typical income over the whole population, or mode-based measurement represents the income of majority workers. The causes of income inequality have been attracting a lot of attention in literature and public. Education and experience are often cited as important factors for income. Gender pay gap has been an issue for many regions, particularly in some developing countries. The difference in wage between public sector staff and private workers changes from country to country and region to region. The mode regression or mode-based regression analysis provides a direct and powerful tool to explore the typical income and the return to education, experience, gender, sector, and so on. Mode-based clustering techniques have also been developed [1].

The mode regression models the relationship between the conditional mode of the dependent variable y^* and covariates x as

$$y^* = x'\beta + \varepsilon \tag{1}$$

where vector $x = (1, x_1, \ldots, x_p)'$, $\beta = (1, \beta_1, \ldots, \beta_p)'$ is the unknown parameter vector, and ε is the model error. Let mode$(y^* \mid x)$ be the mode of y^* conditional on x and then mode$(y^* \mid x) = x'\beta \Leftrightarrow$ mode$(\varepsilon \mid x) = 0$. Usually, one assumes that the density of ε is wider than $[-w, w]$ for a $w > 0$ suitably chosen.

Lee [2] first considers truncated model regression where y^* is truncated from below at c by y or y^* is observed only when $y^* > c$. That is, $y = \max(y^*, c)$. Examples of truncated regression include (1) candidates of students who want to nominate a university prize are required to have a minimum examination mark of 70 out of 100 to qualify for the entry. Thus, the sample is truncated at an examination mark of 70. (2) A researcher has data for a sample of British citizens whose income is above the poverty line. Hence, the lower part of the distribution of income is truncated. Truncated regression cannot be fitted by ordinary least squares (OLS) regression, as OLS regression will not adjust the estimates of the coefficients to take into account the effect of truncation, so that the estimated coefficients may be severely biased. This can be conceptualized as a model specification error [3].

Under model (1), given observations on $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, we aim at estimating β. This is the main task of fitting the model. Lee's [2] mode regression estimates β by

$$\hat{\beta} = \arg\max_{\beta} \left(n^{-1} \sum_{i=1}^{n} I\left[\left| y_i - \max(x'\beta, c + w) \right| \le w \right] \right) \tag{2}$$

where $I[\cdot]$ is the indicator function, which takes the value 1 if the condition inside [ ] is satisfied and 0 otherwise.

But little progress on mode regression has been made due to computation difficulty although some modifications of the proposed mode regression have been developed (e.g., see [4–7]). The difficulty in computing these estimators arises because the objective function consists of indicator function $I[\cdot]$ and involves in an absolute function and maximum operator. So that the objective function in Eq. (2) is neither convex nor differentiable but may result in a large number of local maxima. See further details in Section 2. Therefore, the estimation in Eq. (2) is an NP-hard problem, so that standard optimization algorithms will perform poorly if they tend to get trapped in local maxima, or they may not be applicable if analytical gradients are required, because standard optimization tools, which require the objective function to be differentiable and/or convex, may fail to discover the true mode regression function.

Currently, rather than dealing with the NP-hard problem directly, some attempts were made to solve the computation of mode regression via replacing rectangular kernel in Eq. (2) by a "smooth" version [4, 5]. However, these smooth versions of mode estimators, in spite of their improved asymptotical properties, may not estimate mode but estimate something else. This big issue is often ignored in literature.

For example, the smoothing version of mode estimator of Kemp and Silva [5] is to maximize $Q_n(\beta) = n^{-1}\sum_{i=1}^{n} K_h\left(y_i - x'\beta\right)$ with $K_h = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$ and the standard normal density as kernel function $K(\cdot)$, and then the estimator is closer to mean than to mode due to the quadratic property of normal density function. Also, a careful selection of bandwidth h which requires tending to zero is not available.

Lee [4] employed a quadratic kernel (QME) to smoothing the rectangular kernel and estimated β by

$$\begin{aligned}\hat{\beta} &= \arg\max_{\beta}\left(n^{-1}\sum_{i=1}^{n}\left[w^2 - \{y_i - \max(x'\beta, c+w)\}^2\right] I\left[\left|y_i - \max(x_i'\beta, c+w)\right| \le w\right]\right)\\ &\Leftrightarrow \arg\max_{\beta}\left(n^{-1}\sum_{i=1}^{n}\max\left[w^2 - \{y_i - \max(x'\beta, c+w)\}^2, 0\right]\right)\end{aligned} \tag{3}$$

This QME is quite similar to the symmetrically trimmed least squares (STLSs) estimation in Powell [8], which, instead of maximization, minimizes

$$\hat{\beta} = \arg\min_{\beta} n^{-1}\sum_{i=1}^{n}\left[y_i - \max(0.5y_i + 0.5c, x'\beta)\right]^2 \tag{4}$$

However, QME is quite sensitive to the choice of w whose optimal value is difficult to derive in practice. And the STLS, which strongly depends on the symmetric requirement of y conditional on, needs the symmetry up to $\pm x'\beta$. That is, STLS requires global symmetry if x is unbounded.

The other way to develop efficient algorithms for the truncated mode regression objective function (2) could be the emulation algorithms (EA) [9, 10], which compute the truncated mode regression estimator by checking every critical point and solving maximum score estimation as a nonlinear programming problem. However, EA exhibits a high degree of complexity in its implementation. EA may achieve convergence to local minima, whereas obtaining a global minimum requires a heavy computational load, something that renders its use in solving real problems impractical.

This paper aims to introduce meta-heuristic methods for computing the elegant rectangular mode regression estimator so that one could not only fit the mode regression but also improve computation efficiency. The recommended heuristic method for dealing with complicated objective function with many local optima could be the popular simulated annealing (SA) [11], because SA provides a means to avoid getting stuck in local optima by accepting worse neighbors in hopes of finding a global optimum. However, SA has not been used in mode regression fitting.

The paper is organized as follows. In Section 2, we outline the issue of many local optima of Lee's estimator and SA computation. Section 3 compares different algorithms or estimation methods such as QME, STLS, EA, and SA via (Monte Carlo) simulation study. In Section 4, we fit a multiple mode regression model for the analysis of a real income data via SA algorithm. The final section concludes with brief remarks.

2. Lee's estimator with many local optima and SA algorithm

As mentioned earlier, truncated mode regression fitting problem may be seen as a global optimization problem (GOP) with many local optima. SA is applied to optimize Eq. (2), which represents a nonlinear objective function. Then, the estimation equation can be formulated as follows:

$$f(\beta) = n^{-1}\sum_{i=1}^{n} I\left[\left|y_i - \max(x'\beta, c + w)\right| \le w\right] \tag{5}$$

Note that $f(\beta)$, as a function of regression coefficient β, is an unconstrained nonlinear and non-smooth function; it is difficult to calculate its optimizer. Moreover, $f(\beta)$ is not convex in β. Therefore, an effective tool for solving global optimization problems is required. Let us demonstrate this issue via a simple example and graphical representation of the objective function. Draw an independent sample size of 10 from a standard normal distribution, i.e., $x_i \sim N(0, 1)$, $i = 1, 2, \ldots, 10$, and let the i.i.d error term ε also have a standard normal distribution. So the dependent variable y^* can be generated from the mode regression $y_i^* = \beta x_i + \varepsilon_i$. Then we obtain observations of y_i from y_i^* via truncated point $y_0 = 0$. This gives a 52% heavy truncated on average. The objective function $f(\beta)$ against β's values from this model can take a form as plotted in **Figure 1**. Clearly, the objective function under $w = 1$ has many local maxima, which

http://dx.doi.org/10.5772/66070

lie in the range between −0.5 and 0.25. When w = 1.5, this objective function still has many localities but in the range between −0.75 and 0.25. However, **Figure 1** shows that these localities under w = 0.5 fall in a much narrower range than those under w = 1 and w = 1.5.

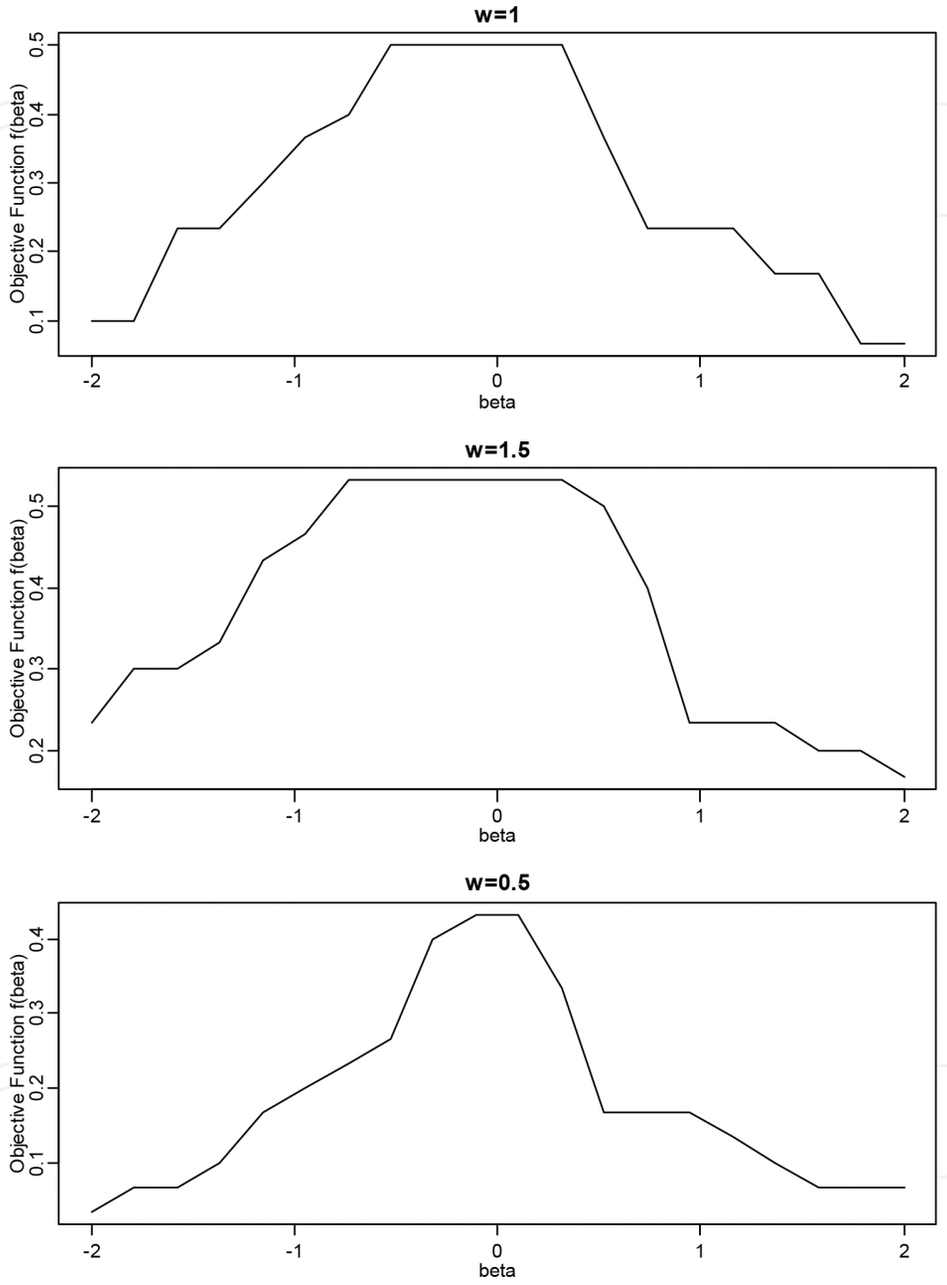


Figure 1. The objective function of a simple example with w = 1, w = 1.5, and w = 0.5.

The SA algorithm proposed by Kirkpatrick, Gelatt, and Vecchi [11], as a local search meta-heuristic, is characterized by an acceptance criterion for neighboring solutions that adapts itself at run time. In this chapter, we process the data by the R software. One of the packages to implement SA in R is *stats* with *sann* as an option of *optim* function [12]. An alternative SA is the *GenSA* function in specific R-package *GenSA*. The SA flow diagram is presented in **Figure 2**.

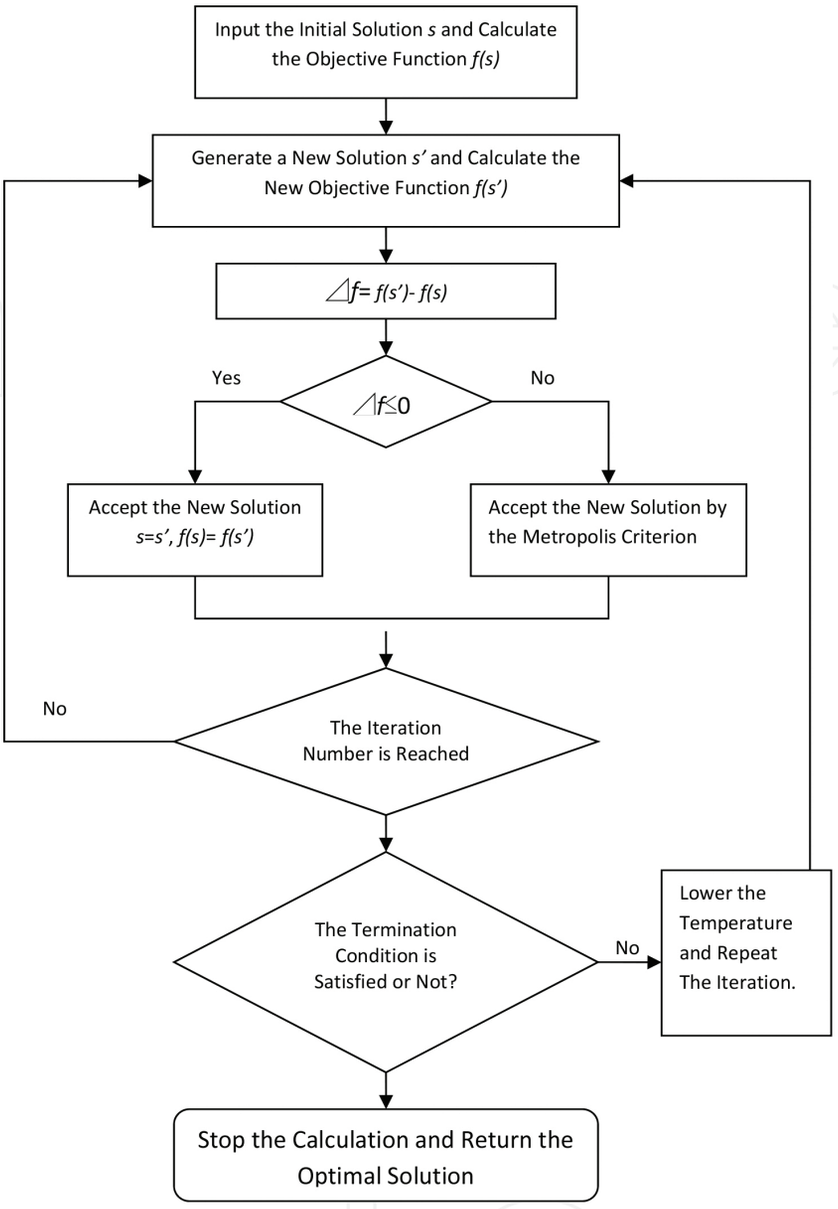


Figure 2. The flow diagram of SA.

3. Numerical comparison

Following Lee ([2, 4]) consider the mode regression model

$$y_i = 0 + x_i + \varepsilon_i, i = 1, 2, \cdots, n \tag{6}$$

where sample size $n = 30$, x_i follows a standard normal random variable, and the model error ε_i is generated from a standard normal distribution.

http://dx.doi.org/10.5772/66070

The four methods—QME (quadratic kernel smoothing method in Ref. [2]), STLS (trimmed least square method in Ref. [8]), emulation algorithm (EA in Ref. [9, 10]), and SA [11]—are implemented for numerical comparison of optimization of $f(\beta)$ in terms of β, where β simply stands for the slope coefficient whose true value is 1.

The data generated from the model are under four different distributions (normal, Cauchy, logistic, and gamma) for model error ε and two different numbers of truncated schemes: 25 and 50%. For each of these five different distributions of ε and each truncated scheme, we implement all four methods to estimate the slope coefficient 200 times, respectively, via simulation.

Then the performance criteria of each method consist of bias of the estimator of slope coefficient, standard deviation (STD), root mean square errors (RMSE), lower quartile (LQ), median (MED), and upper quartile (UQ) of estimation. At each simulation of 200 times, the bias is the difference between estimate and the true value; then the bias we collect for comparison is a simple average of 200 times of replications. The computation of STD, RMSE, LQ, MED, and UQ are based on 200 times of replications.

Table 1 reports the results by different designs which are the combination of truncation rate and distribution of ε.

	BIAS	**SE**	**RMSE**	**LQ**	**MED**	**UQ**
Design 1: 50% truncation, standard normal						
w = 0.5	0.535	1.451	1.563	0.910	1.341	1.811
w = 1.0	0.502	1.011	1.217	0.925	1.244	1.930
w = 1.5	0.523	1.345	3.492	0.696	1.112	1.748
w = 2.0	0.591	1.876	1.729	0.997	1.270	1.840
Design 2: 25% truncation, standard normal						
QME	0.085	0.331	0.345	0.839	1.042	1.269
STLS	0.023	0.210	0.208	0.873	0.989	1.136
EA	0.067	0.023	0.022	0.891	0.993	1.542
SA	0.065	0.381	0.412	0.893	1.114	1.458
Design 3: 50% truncation, standard normal						
QME	0.515	1.139	1.248	0.895	1.213	1.677
STLS	0.200	0.502	0.542	0.913	1.090	1.354
EA	0.206	0.489	0.512	0.807	1.051	1.431
SA	0.215	0.431	0.387	0.801	1.098	1.398
Design 4: 50% truncation, standard Cauchy						
QME	0.470	1.812	1.872	0.851	1.270	1.878

	BIAS	SE	RMSE	LQ	MED	UQ
STLS	0.396	6.390	6.949	0.823	1.109	1.605
EA	0.401	1.978	2.231	0.901	1.123	1.589
SA	0.393	1.759	2.217	0.801	1.012	1.598
Design 5: 50% truncation, standard normal						
QME	0.767	2.338	2.479	0.858	1.488	2.620
STLS	0.494	3.765	3.798	0.776	1.189	1.966
EA	0.506	2.634	2.634	0.884	1.052	1.890
SA	0.509	2.031	3.012	0.765	1.175	1.935
Design 6: 50% truncation, gamma (2, 1) mode						
QME	0.615	3.147	3.209	0.879	1.402	2.710
STLS	0.556	3.403	3.450	1.048	1.474	2.168
EA	0.598	3.479	3.581	0.881	1.057	1.111
SA	0.501	2.549	3.000	0.893	1.231	1.786
Design 7: 50% truncation, gamma (3, 1) mode						
QME	0.688	2.361	2.460	0.771	1.613	2.761
STLS	0.705	3.712	3.778	0.731	1.335	2.486
EA	0.676	2.476	3.247	0.790	0.895	1.431
SA	0.617	3.001	3.223	0.801	1.112	1.524

Table 1. $y_i = 0 + x_i + \varepsilon_i$, $i = 1, 2, \cdots, n$, $x_i \sim N(0, 1)$, $n = 30$, and 200 replications; only the slope is reported.

In Design 1, with 50% truncation and the standard normal distribution as the model error, we first check the effect of choosing w on the performance of SA algorithm. We note that $w = 0.5 - 2$ appears to be the range. So we try SA algorithm corresponding to w = 0.5, 1, 1.5, and 2, respectively. The algorithm gives quite "stable" results with w = 0.5, 1, and 1.5 but provides big variation for w = 2 under different replicates. This fact not only concludes that the selection of w may not be necessarily unique but also indicates that larger w may not have better outcome. So in the real data analysis of Section 4, we suppose that w follows a uniform distribution over the interval of $[w_1, w_2]$, where $w_1 = 0.5$ and $w_2 = 1.5$.

Because of the results from Design 1, we use $w = 1$ (a value between 0.5 and 1.5) in Designs 2 and 3 for SA algorithm but use $w = 0.1$, 0.5, 0.9, and 1.3 to construct a weighted version of QME (WQME) which is due to Lee's [2] suggestion that an average of the estimates for several ws may work along with WQME applied to the reasonable range of w. Actually, WQME sometimes performs better than a specific QME. STLS and EA seem to do well in both Designs 2 and 3. This, however, will change for distributions with thicker tails. Furthermore, STLS and EA get much more estimation variation than SA.

http://dx.doi.org/10.5772/66070

In Designs 4–7, we check the sensitivity to the underlying distribution, particularly to the tail behavior of the distribution. We consider the standard Cauchy, standard logistic, and gamma distributions with 50% truncation.

In Designs 4 and 5, except STLS, all methods perform similar. This is also true in Design 7. STLS method, in spite of the small bias on average, gets big estimation variation for most of cases except the case with a normal distribution. For example, among methods, STLS gives the biggest variance and RMSE for fat-tailed distribution (Cauchy) and skewed distributions (logistic and gamma). So STLS is not very reliable for practical application.

In Designs 4 and 5, EA and SA show slightly better than others in some of cases.

A final comment goes to algorithm speed, measured by CPU time. We found that CPU times for all algorithms used in **Table 1** are comparable, typically in the range between 5 and 10 s.

4. Returns of human capital in China

China has experienced rapid economic growth in the past 30 years, but the increase in wage inequality is gradually a serious problem and should be given more attention. We will analyze the important factors, which affect the incomes for most of Chinese workers. This study consists of an analysis of annual income of 1967 Chinese citizens with ages between 18 and 55 in 2008. The data is provided by Chinese Social Survey Open Database (CSSOD) (http://www.cssod.org/index.php in Chinese). We aim to check how education, experience, sex, district, and job sector affect the typical income. One may use mean regression to carry out the analysis, but average income is largely affected by small number of high-income receivers, so that the resulting analysis does not represent the majority people or the typical case. For example, majority workers often see their income as far lower than the reported average income. Therefore, mode regression is an ideal model to carry out the analysis.

We use a standard log-linear Mincer formulation:

$$\log Y = \beta_1 + \beta_2 Edu + \beta_3 Exp + \beta_4 Exp^2 + \beta_5 Gender + \beta_6 Dis + \beta_7 Sector + \varepsilon \tag{7}$$

where log Y is the logarithm transform of annual income (in 1000 Yuan) and truncated by 2, as majority annual income is more than 8000 Yuan. *Edu* is the number of years of schooling, *Exp* is potential experience (approximated by the age minus years of schooling minus 7), *Gender* is equal to 1 for male and 0 otherwise, *Dis* is equal to 1 for workers from the eastern China and 0 otherwise, *Sector* is equal to 1 for private sector workers and 0 otherwise, and ε is the model error. The estimate coefficient β_i means that a unit increase in the predictor variable results in an increase in $(100(\exp(\beta_i)-1))\%$.

We now carry out the mean and median regression analysis. That is, we fit the model (7) by the mean regression and the median regression model, respectively. By implementing the *lm*

and *rq* (in R-package *quantreg*) function in the R software, we obtain the fitted mean and median results as shown in **Table 2**.

	Mean regression	**Median regression**	**Mode regression**
Intercept	113.884***	134.376***	132.179***
	(9.755)	(11.486)	(228.825)
Education	9.381***	8.501***	3.519***
	(14.418)	(12.614)	(63.185)
Experience	2.103***	1.879***	3.887***
	(3.630)	(3.314)	(56.280)
Experience^2	−0.049***	−0.045***	−0.192***
	(−3.569)	(−3.173)	(−47.839)
Gender	27.919***	25.912***	20.921***
	(8.347)	(7.585)	(232.146)
District	44.463***	43.874***	42.571***
	(13.249)	(12.714)	(447.565)
Sector	7.024*	−2.418	−3.291***
	(1.919)	(0. 520)	(−23.626)

Note: Asymptotic standard errors are in parentheses for the mean and median regression.
* $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$ (two tailed)

Table 2. The results of three different regressions (%).

Table 2 shows that each additional year of education increases the conditional-mean income by a factor of exp(0.09381) = 1.09835, which indicates a 9.835% increase. And the fitted median regression model gives a coefficient of 0.08501, which indicates that one more year of education increases the conditional-median income by exp(0.08501) = 1.08873 or a 8.873% increase. That is, for small values of the estimated coefficient β_i, this is approximately 100 β_i%.

The experience entering the model in quadratic form is due to the most popular version of the Mincer equation [13], which includes a quadratic function in years of potential experience to capture the fact that on-the-job training investments decline over time in a standard life cycle human capital model. The estimated coefficients of years of experience and its square both are significant in the mean and the median regression. That is, the Chinese worker's income shows an inverted U-shaped relationship with years of experience and reaches the maximum at years of 21.459 and 20.878, respectively. This means that experience within about 20 years does make difference on the income but experience longer than 20 years would not add anything more for the conditional-mean or the conditional-median income.

http://dx.doi.org/10.5772/66070

Compared with being female, being male increases the conditional-mean income by 100[exp(0.27919)−1]% = 32.206% according to the mean regression results but by 100[exp(0.25912)−1]% = 29.579% according to the median regression results. In other words, the conditional-mean income for male is 32.206% higher than it is for female, and male's conditional-median income is 29.579% higher than female's, with other covariates held constant.

Similarly, the conditional-mean and conditional-median income difference between workers from the west and the east of China both is about 100[exp(0.44)−1]% = 55.271%.

For the dummy variable sector, the results of the mean regression indicate that the conditional-mean income of workers from private sectors is greater than from public sectors by a factor of exp(0.07024) = 1.072766, that is, a 7.277% increase in conditional-mean income. However, according to the result of the median regression model, the coefficient of sector is not significant, which means that there is no difference between the public and private sectors for the conditional-median income.

As is known that the selection of w may not be unique for the mode regression, we suppose that w ~ U[0.5, 1.5] and draw 300 different random ws from this uniform distribution. Based on the 300 different ws, we implement the *GenSA* function in the R-package *GenSA* to fit the multivariate mode regression and obtain 300 different vectors of the coefficients $\widehat{\beta}$. Following the law of large numbers, the mean of these estimated $\widehat{\beta}_i$ must approach to the real β_i. Then we can use the two-tailed T test for checking if hypotheses about $\widehat{\beta}_i = 0$ are true or not. The mean and the two-tailed T test results of these seven coefficients β_i based on the 300 different mode regressions are shown in **Table 2**.

According to the results of mode regression based on the SA algorithm, the estimated returns to an additional year of education are about 3.519%, and the education has a much smaller effect on the conditional-mode income. Maybe this is one of the reasons why the typical income is lower than the mean and median income.

The coefficients of years of experience and its square both are significant. That is, the relationship between the conditional-mode income and the years of experience is shown in the inverted U-shaped relationship too, but the income for most workers reaches the maximum at 10.122 years of experience. The return to experience, i.e., the derivative of the typical value of log income with respect to experience, is therefore given by a combination of coefficients (3.887 − 2 × 0.192 × experience). The derivative needs to be evaluated at some specified level of experience. Two points were chosen: 5 years of experience, representing fairly new entrants, and 15 years of experience, representing experienced workers. And these two points give the combination of coefficients 1.967 and −1.873. That is, the return to experience of 5 years is 1.967% and experience of 15 years is −1.873%. In contrast with the results of the mean and median regression model, the return to experience of 15 years both is positive. The relationship between income and experience based on these three different regression models is represented in **Figure 3**.



Figure 3. The relationship of income and experience.

As shown in **Figure 3**, we conclude that experience does make a great contribution to the conditional-mode income within 10 years, but experience longer than 10 years would not add anything more for it. That is, the conditional-mode income increases rapidly due to experience and decreases rapidly too after reaching the peak income. However, experience returns are positive until about 20 years to the conditional-mean and conditional-median income. Maybe this is another important reason why the typical income is lower than the mean and median income.

The conditional-mode income from a man is about 20.921% more than from a woman, which means that the gender pay gap is slightly less serious for majority workers compared with the mean and median regression. Similarly, the conditional-mode income from workers of eastern China is about 42.571% more than from worker of western China.

The private sector does not have positive return to mode income; this is in sharp contrast with the results of the mean regression. Concretely speaking, the conditional-mode income for the

public sector worker is 3.291% higher than for the private sector worker. This result is the opposite of what is shown in the mean regression model.

The results from the analyses of both mean and the mode regression models are inconsistent with the economic intuitions. In China, private sector includes private enterprises operated by local Chinese; Sino foreign joint ventures; Hong Kong-, Macao-, and Taiwan-funded enterprises; and foreign-funded enterprises. Workers in the private enterprises which account for the vast majority of the private sectors do not have the "compilation" or sign the labor contracts with these enterprises; thus, their rights and interests cannot be guaranteed and their wage is low. Although the workers in Sino foreign joint ventures and Hong Kong-, Macao-, and Taiwan-funded enterprises do not have the "compilation," they gain an attractive income because of these enterprises' relatively good efficiency. Workers in the foreign enterprises have the highest wages, especially for those CEOs. So the average wage level of workers in the private sector is relatively high. Most workers in the public sectors protected by the "compilation," often sign the labor contracts with the stated own enterprises, tend to gain a higher wage than the workers in the private enterprise, and the wage gap is very relatively small. Thus, the majority of workers have a lower wage in the private sector than in the public sector, but on average, the opposite is the case.

5. Conclusion

While mode regression has been found very useful in practical regression analysis, the main goal of this article is to introduce SA algorithm to fit mode regression models. The most popular mode regression model is introduced by Lee [2]. There is no doubt on the elegance of rectangular mode regression estimator of Lee [2], but the main problem with this estimator lies in computation. While the difficulty in obtaining reliable mode regression coefficients limits the application of the mode regression, we propose the SA algorithm for the mode regression estimation and then compare the proposed SA algorithm with other existing methods. To sum up, SA algorithm for fitting truncated mode regression does not require any liberalization or modification of Lee's estimator and solves the corresponding nonlinear optimization problem more efficiently and robustly as a rule.

As an application of the SA algorithm fitting a real data-based mode regression and the illustration of the sensible interpretation of fitted mode regression coefficients by the algorithm, we apply the mode regression model in income inequality analysis of China and some meaningful conclusions are obtained which are different from the mean regression and the quantile regression.

Acknowledgements

The research is supported in part by the National Sciences Foundation of China (11261048) and National Bureau of Statistics of China (2013LY022).

Author details

Maoxi Tian[1,2], Jian He[1] and Keming Yu[3*]

*Address all correspondence to: Keming.Yu@brunel.ac.uk

1 Department of Statistics, Shihezi University, Xinjiang, China

2 Jinhe Center for Economic Research, Xi'an Jiaotong University, Shaanxi, China

3 Department of Mathematics, Brunel University, London, UK

References

[1] Li J., Ray S., Lindsay B. G. A nonparametric statistical approach to clustering via mode identification. Journal of Machine Learning Research. 2007; 8(4):1687–1723. DOI: 10.1007/s10846-007-9145-xLi

[2] Lee M. J. Mode regression. Journal of Econometrics. 1989; 42(3):337–349. DOI: 10.1016/0304-4076(89)90057-2

[3] Heckman J. J. Sample selection bias as a specification error. Econometrica. 1979; 47(1): 153–161. DOI:10.2307/1912352

[4] Lee M. J. Quadratic mode regression. Journal of Econometrics. 1993; 57:1–19. DOI: 10.1016/0304-4076(93)90056-B

[5] Kemp G. C. R., Silva J. M. C. S. Regression towards the mode. Journal of Econometrics. 2012; 170(1):92–101. DOI:10.1016/j.jeconom.2012.03.002

[6] Yu K., Aristodemou K. Bayesian mode regression. Technical report (2012). arXiv: 1208.0579v1

[7] Yao W., Li L. A new regression model: modal linear regression. Scandinavian Journal of Statistics. 2013; 41(3):656–671. DOI: 10.1111/sjos.12054

[8] Powell J. L. Symmetrically trimmed least squares estimation for Tobit models. Econometrica. 1986; 54(6):1435–1460. DOI:10.2307/1914308

[9] Pinkse C. A. P. On the computation of semiparametric estimates in limited dependent variable models. Journal of Econometrics. 1993; 58(1–2):185–205.

[10] Fitzenberger B. A guide to censored quantile regressions. Handbook of Statistics. 1997; 15(97):405–437. DOI:10.1016/S0169-7161(97)15017-9

[11] Kirkpatrick S., Gelatt C. D., Vecchi M. P. Optimization by simulated annealing. Science. 1983; 220(4598):671–680.

[12] Theussl S., Borchers H. CRAN Task View: Optimization and Mathematical Programming. [Internet]. 2015. Available from: https://cran.r-project.org/web/views/Optimization.html

[13] Mincer J. Schooling, Experience, and Earnings. New York: Columbia University Press; 1974.

IntechOpen

IntechOpen

IntechOpen

IntechOpen