Stat Papers (2015) 56:411–430
DOI 10.1007/s00362-014-0588-4

REGULAR ARTICLE

Empirical likelihood based modal regression

Weihua Zhao · Riquan Zhang · Yukun Liu · Jicai Liu

Received: 24 December 2012 / Revised: 2 March 2014 / Published online: 31 March 2014

Abstract In this paper, we consider how to yield a robust empirical likelihood estimation for regression models. After introducing modal regression, we propose a novel empirical likelihood method based on modal regression estimation equations, which has the merits of both robustness and high inference efficiency compared with the least square based methods. Under some mild conditions, we show that Wilks' theorem of the proposed empirical likelihood approach continues to hold. Advantages of empirical likelihood modal regression as a nonparametric approach are illustrated by constructing confidence intervals/regions. Two simulation studies and a real data analysis confirm our theoretical findings.

Keywords Empirical likelihood · Modal regression · Robust · Confidence region

Mathematics Subject Classifications Primary 62G10, Secondary 62G08

1 Introduction

It is well-known that the ordinary least squares estimator (LSE) is the most efficient estimator of the regression coefficient in linear regression models when the noise follows a normal distribution. However, departure of the error distribution from normality

W. Zhao
School of Science, Nantong University, Nantong 226007, People's Republic of China
e-mail: zhaowhstat@163.com

R. Zhang · Y. Liu (✉) · J. Liu
School of Finance and Statistics, East China Normal University, Shanghai 200241, People's Republic of China
e-mail: ykliu@sfs.ecnu.edu.cn

may severely reduce the efficiency of the LSE, particularly when the errors are heavy-tailed and/or including outliers. One remedy is to remove influential observations from the least-square fit. Another approach, termed robust regression, is to replace the least square loss criterion by outlier-resistant loss criteria in the estimation procedure. Considering that outliers are often genuine data in certain circumstances such as income analysis, procedures like robust regressions, which accommodate rather than directly remove the outliers, will be more efficient.

Suppose we have a simple random sample $\{(y_i, \mathbf{x}_i) : i = 1, 2, \ldots, n\}$ from the following classical linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ and the noise ϵ_i, independent of $\mathbf{x}_i$, are i.i.d. random variables with mean zero. Robust regression estimators, introduced by Huber (1981), were obtained by minimizing $\sum_{i=1}^n \rho(\theta; \mathbf{x}_i)$ with respect to θ, where ρ is a loss function. There are three popular robust regression estimators in the literature with different choices of loss function. The loss criterion $\rho(x) = |x|$ leads to the median regression estimation which is a special case of quantile regression (Koenker and Bassett 1978). The other two choices of $\rho(\cdot)$, i.e. Huber loss and Tukey bisquare loss, corresponds to two Huber's robust estimators (Huber 1981). In particular, if the loss function is chosed to be the log-likelihood function, we obtain the usual maximum likelihood estimator. The above three estimators are also referred as M-type robust estimators. Unfortunately, the median estimator may lose efficiency when there are no outliers or the error distribution is normal; Also it may not be unique since the loss function $\rho(x) = |x|$ is not strictly convex. Huber's robust estimators have high efficiency if an optimal transitional point is available; it is rather difficult to adaptively choose such an optimal transitional point in practice (Rousseeuw and Leroy 1987). Robust regression estimation also gains many developments in recent years, including composite quantile regression (Zou and Yuan 2008), convex combinations of the L_1 and L_2 loss criteria with flexible weights (Chen et al. 2010), rank-based estimation methods (Johnson and Peng 2008), and Modal regression (Yao and Li 2013; Yao et al. 2012).

We have two main goals in this paper, both of which are motivated by Yao and Li's (2013) modal regression. The first goal is to propose a new modal regression after investigating the properties of modal regression estimation method. Yao and Li (2013) showed that the convergence rate of the modal regression coefficient estimator is slower than root-n, where n is the sample size. Under different conditions, we find that this rate can still be root-n if we take the involved bandwidth h as a constant. In doing so, the asymptotical variance of the modal regression coefficient estimator will depend on h, which can be further regarded as a tuning parameter. A data-driven method is provided to estimate the optimal bandwidth which minimizes the asymptotical variance of the modal regression coefficient estimator. Since the resulting estimation procedure has the same form as Yao and Li's (2013) method, we still call it modal regression estimation (MRE), although the two methods are different in essence. Our simulation results indicate that the MRE not only has very good robustness for data sets containing outliers or having a heavy-tail error distribution, but also is as asymptotically efficient

as least-square-based method when there are no outliers or the error distribution follows a normal distribution.

As the second goal of the paper, we propose an empirical likelihood (EL; Owen 1991) based modal regression method to construct confidence regions/intervals or test hypotheses for the regression coefficients. The aforementioned regression methods usually focus on point estimation. Apart from point estimation, confidence regions or intervals of regression coefficients are also important to evaluate the goodness of estimation methods. The EL is an efficient nonparametric likelihood tool that has a number of nice properties (Owen 1988, 1990, 1991). For example, it is flexible in incorporating auxiliary information; the EL ratio statistic usually has a chisquare limiting distribution; and the EL based confidence regions have data-driven shapes, etc. For a more thorough review on EL, we refer the reader to Owen (2001), Chen and Keilegom (2009), Wei et al. (2012), Zi et al. (2012) and references therein. In this paper, we show that the EL ratio based on modal regression estimation equation still follows a chisquare limiting distribution. Given the robustness to outliers of the modal regression and the estimation efficiency of the EL, we expect the resulting EL based modal regression to be robust and efficient when applied to test hypotheses and construct confidence regions. By simulation study, we find that the confidence intervals (regions) based on the proposed method are shorter (smaller) than those based on least square methods when the error follows non-normal distributions.

The rest of the paper is organized as follows. In Sect. 2, we review the modal regression, and study the asymptotical normality of the modal regression estimator taking the bandwidth h as a constant. An adaptive optimal bandwidth is presented for practical purpose. In Sect. 3, we propose the EL based modal regression estimation method for the regression coefficient. A nonparametric Wilks theorem for such an EL ratio statistic is proved. Simulation studies and a real data analysis are provided in Sects. 4 and 5, respectively. Section 6 concludes. For clarity, all technical proofs are deferred in the Appendix.

2 Modal regression

2.1 Modal regression estimation

We begin by briefly reviewing the background and mathematical foundation of modal regression. Mean, median and mode are three important numerical characteristics of distribution. Mode, the most likely value of a distribution, has wide applications in astronomy, biology and finance, where the data is often skewed or contains outliers. Compared with mean, mode has the advantage of robustness, which means that it is resistent to outliers. Moreover, since modal regression focuses on the relationship for the majority of data and summaries the "most likely" conditional values, it can provide more meaningful point prediction and larger coverage probability for prediction than others when the data is skewed or contains outliers.

For model (1), modal regression Yao and Li (2013) estimates the modal regression parameter $\boldsymbol{\beta}$ by maximizing

$$Q_h(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{i=1}^{n}\phi_h\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right), \tag{2}$$

where $\phi_h(t) = h^{-1}\phi(t/h)$, $\phi(t)$ is a kernel density function and h is a bandwidth, determining the degree of robustness of the estimator. As noted by Yao and Li (2013) and Yao et al. (2012), the MRE method usually produces robust estimates due to the nature of mode. When the error distribution is symmetric and has only one mode at the center, then mean regression, median regression and modal regression all estimate the same regression coefficient. For example, we may choose $\phi(t)$ to be the standard normal density function or the Gaussian kernel.

Here is the justification for the claim that the object function (2) can be used to estimate the modal regression. Consider the case that only the intercept $\beta = \beta_c$ is involved in linear regression (1). Then the object function $Q_h(\boldsymbol{\beta})$ defined in (2) reduces to

$$Q_h(\beta_c) \equiv \frac{1}{n}\sum_{i=1}^{n}\phi_h\left(y_i - \beta_c\right). \tag{3}$$

which can be regarded as a kernel estimate of the density function of y at $y = \beta_c$. Therefore, the maximizer of (2) is the mode of the kernel density function based on $y_1, \ldots, y_n$. As $n \to \infty$ and $h \to 0$, the mode of kernel density function will converge to the mode of the distribution of y under certain conditions (Parzen 1962).

In contrast to other estimation methods, modal regression treats $-\phi_h(\cdot)$ as a loss function, which is a special M-type robust regression mentioned in Sect. 1. Since modal regression can estimate the "most likely" conditional values, it can provide more robust and efficient estimation than other existing methods. Lee (1989) used the uniform kernel and Epanechnikov kernel for $\phi(\cdot)$ to estimate the modal regression, respectively. However, their estimators are of little practical use because the object function is non-differentiable and its distribution is intractable. Scott (1992) mentioned the modal regression, but little methodology is given on how to implement it in practice. Recently, Yao and Li (2013) suggested using the Gaussian kernel for $\phi(\cdot)$ and developed MEM algorithm to compute modal estimators for linear models. Yao et al. (2012) investigated the estimation problem in nonparametric regression using the method of modal regression, and obtained a robust and efficient estimator for the nonparametric regression function. Their estimation procedure is very convenient to implement for practitioners and the result is encouraging for many non-normal error distributions. In addition, Yu and Aristodemou (2012) studied modal regression from Bayesian perspective.

2.2 Theoretical property

In this subsection, we first take the bandwidth as a constant and establish the asymptotical normality of the proposed modal regression estimator (MRE). The limiting variance of the MRE is found dependent of h. We recommend an optimal bandwidth by minimizing the limiting variance.

The desirable property of the MRE estimator is achieved under certain assumptions on both the error and the kernel function ϕ. Here we assume that the errors ϵ_i's in model (1) are independent and identically distributed (iid), and that the underlying kernel function $\phi(\cdot)$ together with the error distribution satisfies

(C1) $\mathrm{E}(\phi_h'(\epsilon)) = 0$, $F(h) \equiv \mathrm{E}(\phi_h''(\epsilon)) < 0$ and $G(h) \equiv \mathrm{E}(\phi_h'(\epsilon)^2)$ is finite for any $h > 0$;
(C2) There exists $c > 0$ such that $\mathrm{E}\{\rho_{h,c}(\epsilon)\} < \infty$, where $\rho_{h,c}(\epsilon) = \sup_{y:|y-\epsilon|<c} |\phi_h'''(y)|$.

Remark 1 Assumption (C1) is a general assumption for modal regression. See Yao and Li (2013) and Yao et al. (2012). Condition (C2) is used to control the magnitude of the remainder in a third-order Taylor expansion of $Q_h(\boldsymbol{\beta})$. See Eq. (18). The condition $F(h) < 0$ ensures that there exists a local maximizer of $Q_h(\boldsymbol{\beta})$, while the condition $\mathrm{E}\{\phi_h'(\epsilon)\} = 0$ guarantees the consistency of this local maximizer, the proposed estimator of $\boldsymbol{\beta}$. Conditions (C1) and (C2) are satisfied if both the error density function and $\phi(\cdot)$ are symmetric and the error has a unique mode. More specifically, when conditions (C1) and (C2) hold, the estimated function based on modal regression is generally the same for mean regression, although the MRE are more robust to outliers. In applications, these conditions will roughly be satisfied if the residual histogram of our modal regression is roughly hell-shaped or has only one mode. We may first apply the MRE method, and then check whether the residuals have this property.

Theorem 1 *Suppose* $\{(y_i, \boldsymbol{x}_i) : i = 1, 2, \ldots, n\}$ *are iid observations from model* (1) *where* $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, *the error* ϵ_i *and the covariate* $\mathbf{x}_i$ *are independent, and* $(\epsilon_i, \mathbf{x}_i^\tau)$*'s are iid with finite covariance matrix. For fixed bandwidth* $h > 0$, *if the error distribution and* ϕ *satisfy conditions (C1) and (C2), then there exists a local maximizer* $\hat{\boldsymbol{\beta}}$ *of* $Q_h(\boldsymbol{\beta})$ *in* (2) *such that* $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{\Omega})$, *where* $\xrightarrow{d}$ *stands for convergence in distribution and* $\boldsymbol{\Omega} = \{G(h)/F^2(h)\}\boldsymbol{\Sigma}^{-1}$ *with* $\Sigma = Cov(\mathbf{x}_i)$ *positive definite.*

A Proof of Theorem 1 is given in the Appendix. If $\mathrm{Var}(\epsilon_i) = \sigma^2$, the asymptotic variance of the least square estimator (LSE) is equal to $\sigma^2\boldsymbol{\Sigma}^{-1}$. This together with Theorem 1 implies that the asymptotic relative efficiency of the MRE over the LSE is $r(h) = \sigma^2 F^2(h)/G(h)$. Theoretically, the larger the asymptotic relative efficiency is, the better the former estimator is. If we take h as a tuning parameter for choosing a good MRE, an ideal choice of this tuing parameter is

$$h_{\mathrm{opt}} = \arg\max_h r(h) = \arg\max_h F^2(h)/G(h). \tag{4}$$

This bandwidth gives the best MRE estimator compared with the LSE from the viewpoint of asymptotical variance. A distinct property of h_{opt} from the usual bandwidth in nonparameter regression is that this h_{opt} depends not on the sample size n but only on the error distribution and the first two derivatives of ϕ.

2.3 Bandwidth selection

Bandwidth plays an important role in order to obtain the robust estimation. We provide a bandwidth selection method for the practical use of the MRE. Following the idea of Yao et al. (2012), we first estimate $F(h)$ and $G(h)$ by

$$\hat{F}(h) = \frac{1}{n}\sum_{i=1}^{n} \phi_h''(\hat{\epsilon}_i) \quad \text{and} \quad \hat{G}(h) = \frac{1}{n}\sum_{i=1}^{n} \left\{\phi_h'(\hat{\epsilon}_i)\right\}^2, \tag{5}$$

respectively, where $\hat{\epsilon}_i = y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ is estimated based on some robust pilot estimates, such as the lease absolute deviation (LAD) estimator or the rank-based estimator (Johnson and Peng 2008). We recommend choosing the bandwidth to be

$$\tilde{h}_{\text{opt}} = \arg\min_h \{\hat{F}(h)\}^2/\hat{G}(h). \tag{6}$$

A quick method of solving this minimization problem is the grid search method. As done by Yao et al. (2012), we may choose the possible grids points to be $h = 0.5\hat{\sigma} \times 1.02^j$ $(0 \le j \le k)$ for $k = 50$ or 100, where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$.

3 Empirical likelihood based modal regression

In this section, we propose empirical likelihood based modal regression to construct confidence regions for the regression coefficients.

From (2), we can define an auxiliary random vectors (Qin and Lawless 1994)

$$\xi_i(\boldsymbol{\beta}) = \boldsymbol{x}_i\phi_h'(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}), \quad i = 1, \ldots, n. \tag{7}$$

Note that $\mathrm{E}\{\xi_i(\boldsymbol{\beta}_0)\} = 0$ where $\boldsymbol{\beta}_0$ is the true parameter value. According to the empirical likelihood principle, we define the empirical likelihood ratio function of $\boldsymbol{\beta}$ to be

$$\mathcal{L}_n(\boldsymbol{\beta}) = \sup\left\{\prod_{i=1}^{n}(np_i)\Big| p_i \ge 0,\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i\xi_i(\boldsymbol{\beta}) = 0\right\}. \tag{8}$$

Given $\boldsymbol{\beta}$, if $\{(p_1, \ldots, p_n) : p_i \ge 0,\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i\xi_i(\boldsymbol{\beta}) = 0\}$ is an empty set, the likelihood ratio $\mathcal{L}_n(\boldsymbol{\beta})$ will have no definition. In this situation, Chen et al.'s (2008) adjusted empirical likelihood is likely the most straightforward and natural remedy to this dilemma, although the convention defines $\mathcal{L}_n(\boldsymbol{\beta})$ to be zero.

Otherwise $\mathcal{L}_n(\boldsymbol{\beta})$ is well-defined and can be re-expressed as

$$\mathcal{L}_n(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left\{1 + \boldsymbol{\lambda}_\beta^T\xi_i(\boldsymbol{\beta})\right\}^{-1}, \tag{9}$$

where $\boldsymbol{\lambda}_\beta$ is the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i(\boldsymbol{\beta})}{1+\boldsymbol{\lambda}_\beta^T\xi_i(\boldsymbol{\beta})}=0. \tag{10}$$

Accordingly the empirical log-likelihood ratio function is defined as

$$l_n(\boldsymbol{\beta}) =: \log\{\mathcal{L}_n(\boldsymbol{\beta})\} = -\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}_\beta^T\xi_i(\boldsymbol{\beta})\right\}. \tag{11}$$

A feasible and efficient algorithm is needed for the computation of $l_n(\boldsymbol{\beta})$ if one intends to apply the empirical likelihood method. The convex duality method given in Owen (2001, pp. 60–63) can serve this purpose, and it is also adopted in our simulation study.

As expected, we find that when $\boldsymbol{\beta}$ takes its true value $\boldsymbol{\beta}_0$, the empirical log-likelihood ratio $-2l_n(\boldsymbol{\beta}_0)$ still follows a limiting chi-square distribution. This result is summarized in the following theorem.

Theorem 2 *Assume the same conditions as Theorem 1. As n tends to infinity, we have*

$$-2l_n(\boldsymbol{\beta}_0)\xrightarrow{d}\chi_p^2, \tag{12}$$

where χ_p^2 is the chi-square distribution with p degrees of freedom.

According to Theorem 2, the empirical likelihood ratio $-2l_n(\boldsymbol{\beta}_0)$ is asymptotically pivotal; it can be used not only to test the hypothesis $H_0:\boldsymbol{\beta}=\boldsymbol{\beta}_0$, but also to construct confidence regions for $\boldsymbol{\beta}$. Specifically, a modal-regression-empirical-likelihood (MREL) based confidence region with confidence level $(1-\alpha)$ is given by

$$\mathcal{C}_{\mathrm{MREL}}(\boldsymbol{\beta})=\left\{\boldsymbol{\beta}:-2l_n(\boldsymbol{\beta})\le\chi_{p,1-\alpha}^2\right\},$$

where $\chi_{p,1-\alpha}^2$ is the $(1-\alpha)$-quantile of the χ_p^2 distribution. Theorem 2 implies that $\mathcal{C}_{\mathrm{MREL}}(\boldsymbol{\beta})$ constitutes a confidence region for $\boldsymbol{\beta}$ with asymptotically correct coverage probability $1-\alpha$.

4 Simulation study

In this section, we provide simulation results to study the finite-sample properties of the proposed MRE and MREL methods and compare them with existing methods. The proposed MREL method is convenient to be used for confidence interval/region construction, while it reduces to the MRE method when point estimation of the regression coefficient is of interest and the bandwidth is fixed.

We generated data-sets from two models, under which point estimation and interval/region estimation are the respective focuses. Simulation results are computed based on 1000 random samples with the sample size being 50, 100 and 150, respectively. Confidence level is set to be 95 % when confidence interval/region is of interest.

4.1 Example 1

The main goal of this example is to examine the robustness and efficiency of the proposed modal regression estimator (MRE). Let the true regression model be

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \epsilon_i, \quad i = 1, \ldots, n,$$

where the covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^T$ follows a three-dimensional normal distribution $N(0, \Sigma)$ with unit marginal variance and correlation 0.5. The true value of the regression coefficient is $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_3)^T = (1.5, 2, -1.2, 0)^T$. The error ϵ_i is independent of $\boldsymbol{x}_i$. We consider six different error distributions: (1) standard normal distribution, $N(0, 1)$; (2) t-distribution with degree of freedom 3, $t(3)$; (3) standard Laplace distribution, $Lp(0, 1)$; (4) mixture of two normal distributions, $0.9N(0, 1) + 0.1N(0, 10^2)$; (5) mixture of normal-$\chi^2(5)$ distribution, $0.9N(0, 1) + 0.1\chi^2(5)$; (6) mixture of three normal distributions, $0.8N(0, 1) + 0.1N(-10, 1) + 0.1N(10, 1)$. Throughout the paper, we choose and recommend the kernel function ϕ to be the standard normal density function in our MRE method. It can be verified that the conditions of Theorem 1 are all satisfied by all the above error distributions except case (5). We include case (5) in our simulation to investigate the robustness of the proposed MRE method.

For illustration and comparison, we also take the following methods into consideration: least square estimate (LSE), the least absolute deviance estimate (LAD), the composite quantile regression with 9 quantiles (CQR, Zou and Yuan 2008) and the rank regression estimate (RRE, Johnson and Peng 2008). For each method, we report the mean square error (MSE) of the estimate $\hat{\boldsymbol{\beta}}$, i.e., $\text{MSE} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/p$. In order to evaluate the prediction performance of the fitted model, we generated a test sample, e.g. $\{(y_i^{\text{test}}, \boldsymbol{x}_i^{\text{test}}) : i = 1, \ldots, 200\}$, in each simulation, and computed the mean absolute prediction error (MAPE), $\sum_{i=1}^{200} |y_i^{\text{test}} - \hat{y}_i^{\text{test}}|/200$ with $\hat{y}_i^{\text{test}} = (\boldsymbol{x}_i^{\text{test}})^T\hat{\boldsymbol{\beta}}$. The mean and standard error of MSE and MAPE over 1000 replications are reported in Table 1.

From Table 1, we have the following observations. For a given error distribution, the performances of MRE become better and better when the sample size increases. In the case of normal error, as long as the sample size is not too small, the MRE is better than other three robust methods, and it seems to perform almost as well as LSE. And for the Laplace distribution, it is well known that the LAD is the best estimator, nevertheless, the performance of MRE is very close to LAD. For the other four error distributions, it is obvious that MRE outperforms the rest four methods even in case (5) where the conditions in Theorem 1 are not satisfied.

Furthermore, it is worth mentioning that the performances of MRE are significantly better than the others for the three mixture error distributions. Here is a possible reason for this observation. The mixtures can be viewed as populations containing outliers. When data contains severely departed outliers, the modal regression puts more weight on the "most likely" data around the true value, which leads to robustness and efficiency of the proposed MRE.

Table 1 Mean and standard error of MSE and MAPE

n	Method	MSE	MAPE	MSE	MAPE
		$N(0, 1)$		$t(3)$	
50	LSE	**0.0300** (0.0246)	**0.8314** (0.0507)	0.0959 (0.1541)	1.1942 (0.1378)
	LAD	0.0468 (0.0392)	0.8498 (0.0595)	0.0619 (0.0575)	1.1671 (0.1106)
	MRE	0.0347 (0.0312)	0.8371 (0.0572)	0.0565 (0.0532)	1.1617 (0.1087)
	CQR	0.0321 (0.0268)	0.8336 (0.0513)	0.0550 (0.0483)	1.1608 (0.1068)
	RRE	0.0352 (0.0282)	0.8393 (0.0522)	**0.0502** (0.0406)	**1.1574** (0.1035)
100	LSE	**0.0147** (0.0122)	**0.8159** (0.0454)	0.0432 (0.0630)	1.1388 (0.1022)
	LAD	0.0220 (0.0180)	0.8240 (0.0481)	0.0278 (0.0238)	1.1242 (0.0956)
	MRE	0.0155 (0.0128)	0.8167 (0.0455)	**0.0236** (0.0198)	**1.1201** (0.0942)
	CQR	0.0155 (0.0124)	0.8168 (0.0456)	0.0239 (0.0199)	1.1207 (0.0939)
	RRE	0.0168 (0.0136)	0.8175 (0.0454)	0.0237 (0.0198)	1.1274 (0.0947)
150	LSE	**0.0092** (0.0072)	**0.8101** (0.0429)	0.0302 (0.0492)	1.1359 (0.1025)
	LAD	0.0144 (0.0108)	0.8162 (0.0444)	0.0184 (0.0148)	1.1243 (0.0974)
	MRE	0.0096 (0.0076)	0.8105 (0.0432)	**0.0161** (0.0131)	**1.1216** (0.0969)
	CQR	0.0097 (0.0077)	0.8106 (0.0431)	0.0162 (0.0129)	1.1220 (0.0968)
	RRE	0.0108 (0.0085)	0.8115 (0.0435)	0.0158 (0.0128)	1.1197 (0.0945)
		$Lp(0, 1)$		$0.9N(0, 1) + 0.1N(0, 10^2)$	
50	LSE	0.0628 (0.0538)	1.0742 (0.0837)	0.3400 (0.4176)	1.7893 (0.3138)
	LAD	0.0474 (0.0443)	1.0592 (0.0816)	0.0576 (0.0522)	1.5740 (0.2135)
	MRE	0.0505 (0.0479)	1.0624 (0.0826)	**0.0404** (0.0362)	**1.5574** (0.2109)
	CQR	0.0476 (0.0392)	1.0596 (0.0792)	0.0520 (0.0473)	1.5705 (0.2137)
	RRE	**0.0454** (0.0405)	**1.0545** (0.0792)	0.0517 (0.0518)	1.5638 (0.2120)
100	LSE	0.0295 (0.0261)	1.0392 (0.0736)	0.1627 (0.1616)	1.6531 (0.2322)
	LAD	**0.0196** (0.0185)	**1.0251** (0.0707)	0.0271 (0.0224)	1.5331 (0.2032)
	MRE	0.0217 (0.0209)	1.0301 (0.0714)	**0.0179** (0.0156)	**1.5233** (0.2024)
	CQR	0.0212 (0.0185)	1.0300 (0.0713)	0.0226 (0.0190)	1.5281 (0.2029)
	RRE	0.0203 (0.0182)	1.0253 (0.0749)	0.0226 (0.0191)	1.5333 (0.2083)
150	LSE	0.0188 (0.0159)	1.0260 (0.0720)	0.1009 (0.0965)	1.6130 (0.2232)
	LAD	**0.0122** (0.0108)	**1.0177** (0.0711)	0.0179 (0.0151)	1.5347 (0.2116)
	MRE	0.0137 (0.0117)	1.0198 (0.0715)	**0.0121** (0.0102)	**1.5289** (0.2116)
	CQR	0.0136 (0.0114)	1.0200 (0.0712)	0.0146 (0.0122)	1.5317 (0.2114)
	RRE	0.0127 (0.0109)	1.0178 (0.0707)	0.0145 (0.0114)	1.5344 (0.1946)
		$0.9N(0, 1) + 0.1\chi^2(5)$		$0.8N(0, 1) + 0.1N(-10, 1) + 0.1N(10, 1)$	
50	LSE	0.1825 (0.1705)	1.3571 (0.1660)	0.6410 (0.5468)	3.0728 (0.3626)
	LAD	0.0637 (0.0560)	1.2695 (0.1282)	0.0860 (0.0914)	2.7239 (0.2702)
	MRE	**0.0466** (0.0428)	**1.2573** (0.1258)	**0.0447** (0.0914)	**2.6897** (0.2662)
	CQR	0.0710 (0.0643)	1.2698 (0.1280)	0.2270 (0.2327)	2.8660 (0.3274)
	RRE	0.0541 (0.0460)	1.2576 (0.1264)	0.0947 (0.1234)	2.7103 (0.2674)
100	LSE	0.1205 (0.0874)	1.3054 (0.1340)	0.3044 (0.2481)	2.8884 (0.2974)
	LAD	0.0333 (0.0255)	1.2382 (0.1217)	0.0373 (0.0323)	2.6916 (0.2638)

Table 1 continued

n	Method	MSE	MAPE	MSE	MAPE
		$0.9N(0,1)+0.1\chi^2(5)$		$0.8N(0,1)+0.1N(-10,1)+0.1N(10,1)$	
	MRE	**0.0217** (0.0171)	**1.2317** (0.1225)	**0.0192** (0.0160)	**2.6759** (0.2636)
	CQR	0.0412 (0.0330)	1.2397 (0.1208)	0.1342 (0.1153)	2.8025 (0.2889)
50	LSE	0.1825 (0.1705)	1.3571 (0.1660)	0.6410 (0.5468)	3.0728 (0.3626)
	LAD	0.0637 (0.0560)	1.2695 (0.1282)	0.0860 (0.0914)	2.7239 (0.2702)
	MRE	**0.0466** (0.0428)	**1.2573** (0.1258)	**0.0447** (0.0914)	**2.6897** (0.2662)
	CQR	0.0710 (0.0643)	1.2698 (0.1280)	0.2270 (0.2327)	2.8660 (0.3274)
	RRE	0.0541 (0.0460)	1.2576 (0.1264)	0.0947 (0.1234)	2.7103 (0.2674)
100	LSE	0.1205 (0.0874)	1.3054 (0.1340)	0.3044 (0.2481)	2.8884 (0.2974)
	LAD	0.0333 (0.0255)	1.2382 (0.1217)	0.0373 (0.0323)	2.6916 (0.2638)
	MRE	**0.0217** (0.0171)	**1.2317** (0.1225)	**0.0192** (0.0160)	**2.6759** (0.2636)
	CQR	0.0412 (0.0330)	1.2397 (0.1208)	0.1342 (0.1153)	2.8025 (0.2889)
	RRE	0.0283 (0.0228)	1.2396 (0.1232)	0.0384 (0.0368)	2.6793 (0.2591)
150	LSE	0.1009 (0.0588)	1.2873 (0.1255)	0.2031 (0.1637)	2.8025 (0.2770)
	LAD	0.0243 (0.0185)	1.2259 (0.1214)	0.0245 (0.0200)	2.6638 (0.2620)
	MRE	**0.0155** (0.0121)	**1.2213** (0.1215)	**0.0122** (0.0102)	**2.6527** (0.2621)
	CQR	0.0322 (0.0228)	1.2285 (0.1198)	0.1137 (0.0956)	2.7669 (0.2870)
	RRE	0.0204 (0.0156)	1.2310 (0.1208)	0.0246 (0.0218)	2.6657 (0.2720)

The bold numbers correspond to the smallest value for each model setup in terms of MSE or MAPE

Overall, the performance of MRE is desirable and its efficiency gain is more prominent when the data set contains outliers.

4.2 Example 2

We now study the performance of the MREL confidence interval/regions. The usual normality-based least square method (LS) and least square based empirical likelihood method (LSEL; Owen 1991) are also taken into consideration for comparison.

Consider the following model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + 0.5\epsilon_i,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2)^T = (2, 1)^T$. The covariates (x_{i1}, x_{i2}) follows a bivariate normal distribution with mean zero. Both x_{i1} and x_{i2} have univariate variance and their correlation coefficient is 0.8. We generated errors from four distributions: $N(0, 1)$, $t(3)$, $Lp(0, 1)$ and $0.9N(0, 1) + 0.1N(0, 10^2)$. The simulation results are summarized in Table 2 and Fig. 1.

Remark 2 When only β_1 is of interest, the MREL confidence interval of β_1 can be constructed through the profile empirical log-likelihood function $l_n(\beta_1) = \sup_{\beta_2} l_n(\beta_1, \beta_2)$. Similar to usual parametric likelihood, if β_{10} is the true value of

Table 2 Simulated coverage probabilities (CP) of confidence intervals (regions) for β_1, β_2, and $(\beta_1, \beta_2)^T$ and the average lengths (AL) of confidence intervals from three different approaches at nominal level 0.95, where LS denotes the confidence intervals (regions) obtained using least square normal asymptotic method

n	Parameter	Method	LS	LSEL	MREL	LS	LSEL	MREL
			$N(0, 1)$			$t(3)$		
50	β_1	CP	0.9480	0.9290	0.8900	0.9430	0.8980	0.8640
		AL	0.4739	0.4558	0.4942	0.7665	0.7489	0.5894
	β_2	CP	0.9550	0.9360	0.9000	0.9460	0.9080	0.8890
		AL	0.4736	0.4577	0.4865	0.7649	0.7450	0.5743
	(β_1, β_2)	CP	0.9430	0.9150	0.8950	0.9410	0.8790	0.9030
100	β_1	CP	0.9390	0.9350	0.9240	0.9390	0.9250	0.9180
		AL	0.3277	0.3235	0.3307	0.5460	0.5580	0.4109
	β_2	CP	0.9450	0.9360	0.9220	0.9430	0.9230	0.9250
		AL	0.3276	0.3238	0.3278	0.5447	0.5606	0.4123
	(β_1, β_2)	CP	0.9470	0.9440	0.9230	0.9400	0.9070	0.9130
150	β_1	CP	0.9510	0.9420	0.9360	0.9460	0.9310	0.9420
		AL	0.2682	0.2682	0.2709	0.4494	0.4634	0.3368
	β_2	CP	0.9430	0.9410	0.9430	0.9460	0.9330	0.9330
		AL	0.2680	0.2686	0.2709	0.4500	0.4603	0.3379
	(β_1, β_2)	CP	0.9480	0.9460	0.9470	0.9430	0.9100	0.9430
			$Lp(0, 1)$			$0.9N(0, 1) + 0.1N(0, 10^2)$		
50	β_1	CP	0.9370	0.8540	0.9490	0.9530	0.8680	0.9370
		AL	1.8198	1.7735	0.6593	1.7744	1.7155	0.6603
	β_2	CP	0.9410	0.8420	0.9450	0.9510	0.8650	0.9400
		AL	1.8332	1.7833	0.6649	1.7714	1.7242	0.6487
	(β_1, β_2)	CP	0.9220	0.7840	0.9250	0.9480	0.7930	0.9360
100	β_1	CP	0.9410	0.8810	0.9470	0.9490	0.8900	0.9590
		AL	1.2687	1.3014	0.4467	1.2817	1.3198	0.4517
	β_2	CP	0.9430	0.8830	0.9490	0.9560	0.8890	0.9350
		AL	1.2714	1.3209	0.4498	1.2780	1.3287	0.4443
	(β_1, β_2)	CP	0.9490	0.8050	0.9390	0.9410	0.8610	0.9370
150	β_1	CP	0.9480	0.9250	0.9600	0.9510	0.8990	0.9390
		AL	1.0446	1.0932	0.3659	1.0384	1.0698	0.3616
	β_2	CP	0.9350	0.8990	0.9490	0.9410	0.9020	0.9500
		AL	1.0488	1.1197	0.3631	1.0400	1.0924	0.3624
	(β_1, β_2)	CP	0.9510	0.8550	0.9490	0.9490	0.8850	0.9390

β_1, then $-2l_n(\beta_{10})$ has a χ_1^2 limiting distribution as $n \to \infty$. Accordingly a natural MREL confidence interval is given by

$$\mathcal{C}_{\text{MREL}}(\beta_1) = \left\{\beta_1 : -2l_n(\beta_1) \leq \chi^2_{1,1-\alpha}\right\}.$$

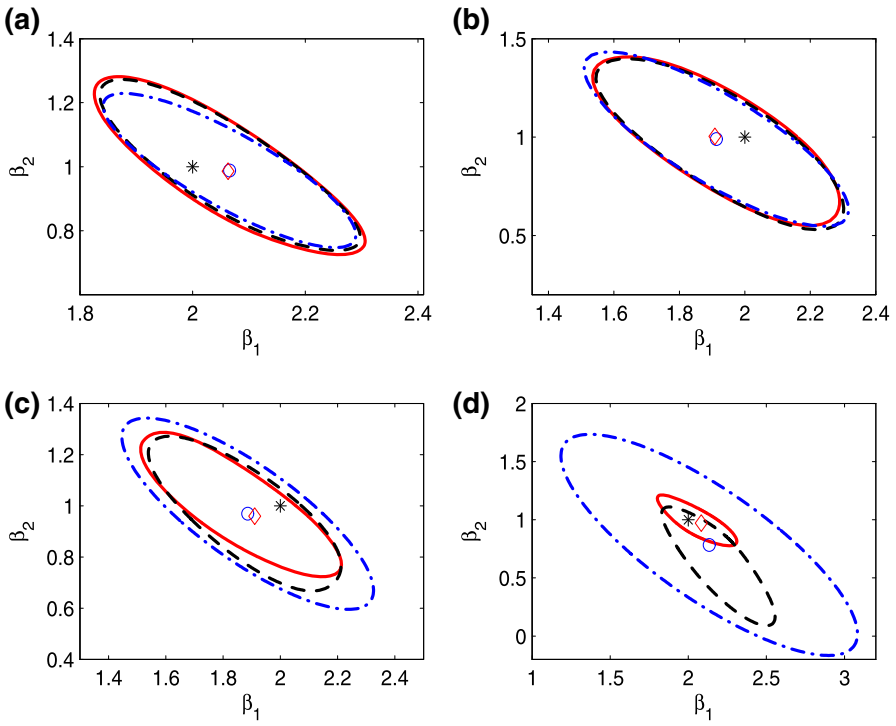

Fig. 1 95 % confidence regions for three different methods in one simulation when sample size $n = 50$: LS (*blue dot dash line*); LSEL (*black dash line*); MREL (*red solid line*), where *asterisk* stands for true value of $(\beta_1, \beta_2)^T$, *circle* and *diamond* denotes least square estimate and modal regression estimate, respectively. (Color figure online)

The construction of $\mathcal{C}_{\mathrm{MREL}}(\beta_2)$ is similar. In Table 2, we also report the marginal confidence interval $\mathcal{C}_{\mathrm{MREL}}(\beta_1)$ and $\mathcal{C}_{\mathrm{MREL}}(\beta_2)$

For a given error distribution, we see that the coverage probability of MREL gets closer and closer to the nominal level as n increases; meanwhile the average lengths of confidence intervals for single parameter become shorter and shorter.

In the case of normal error, the differences among the three methods are small. In particular, the performance of MREL is as well as the least square based methods when the sample size n is large. For the case of non-normal distributions, the average lengths of confidence intervals (regions) for MREL are obviously shorter (smaller) than the other two. It is worth mentioning that the interval length of MREL is only about one third to that of the LS and LSEL when the error follows a mixture normal distribution.

In addition, the coverage probability of LSEL deviates significantly from the nominal level for the three non-normal error distributions when the sample size is small, and it grows very slowly as the sample size increases.

In summary, the MREL has priority over the LS and LSEL methods when the sample size is large in terms of both coverage probability and interval length or region volume. The coverage precision of the MREL confidence interval/region needs improving

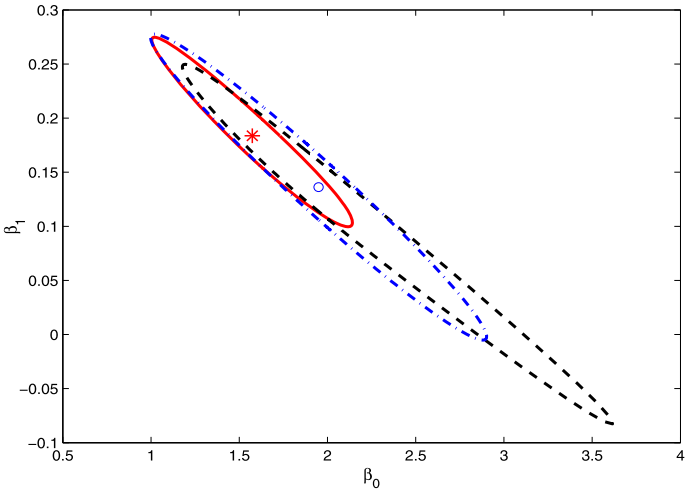

Fig. 2 95 % confidence regions for three different methods: LS (*blue dot dash line*); LSEL (*black dash line*); MREL (*red solid line*), where *circle* and *asterisk* denote the point estimates of LSE and MRE, respectively. (Color figure online)

particularly in the case of small sample sizes. The adjusted empirical likelihood of Chen et al. (2008) and Liu and Chen (2010) or the bootstrap method can serve this purpose

Remark 3 We take the MREL confidence regions in Fig. 2 for example, to illustrate how we computed the confidence boundary given a data-set. The first step is to compute the center, the MRE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Then along any line through the center, two points meeting with the confidence boundary can be found. All points on the confidence boundary will be obtained after we work for all lines. This can be done conveniently in polar coordinate. It is clear that theoretically this method applies to confidence regions of any dimension.

5 Real data analysis

In this section, we apply the proposed method to the analysis of the Education Expenditure Data (Chatterjee and Price 1977). This data set consists of 50 observations from 50 states, one for each state. It has been analyzed by Yao et al. (2012) using nonparametric modal regression. We take the per capita expenditure on public education in a state as the response variable y_i, and take the number of residents per thousand residing in urban areas in 1970 as covariate x_i. And we consider fitting the data by the following linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, 50. \tag{13}$$

In this example, an obvious outlier is that from Hawaii with a very high per capita expenditure on public education compared with other states. The confidence intervals

Table 3 95 % interval estimates for education expenditure data

Method	LS	LSEL	MREL
β_0	(1.1870, 2.7125)	(1.3098, 3.1904)	(1.1246, 2.0327)
β_1	(0.0230, 0.2495)	(−0.0268, 0.2291)	(0.1156, 0.2551)

Table 4 Coverage probability of the confidence region (interval) based on 2000 bootstrap resampling

Method	LS	LSEL	MREL
(β_0, β_1)	0.9485	0.9010	0.9120
β_0	0.9575	0.9635	0.9105
β_1	0.9440	0.9495	0.9220

for β_0 and β_1 respectively based on the LS, LSEL and MREL methods were computed and presented in Table 3. The confidence regions based on the three methods are displayed in Fig. 2. (Here, to alleviate the magnitude difference between the two estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using the original data, we divide both response variable y_i and covariate x_i by 100 for each observation, and then use (13) to fit the transformed data.)

As we can clearly see from Table 3 and Fig. 2, the confidence interval (region) obtained by MREL is shorter (smaller) than the least square based methods, which show that the confidence region obtained by modal regression empirical likelihood not only has the advantage of data-driven nonparametric approach but also is robust to outliers.

To further test the credibility of the confidence region (interval), we also calculated the coverage probability (given in Table 4) of the confidence region/interval based on 2000 bootstrap resamples. As we can see from Table 4, compared with the nominal coverage 95 %, both the two empirical likelihood based methods is not that satisfactory. The bootstrap method and the adjusted empirical likelihood mentioned in Sect. 4.2 can be used to improve the coverage precision.

The comparison of confidence region volume is not fair if the confidence regions under comparison have rather different coverage probabilities. For fair comparison, we calibrate the LS, LSEL and MREL with not their limiting distributions but the empirical distributions based on the 2000 bootstrap statistics. Take the MREL for example. Let $\hat{\boldsymbol{\beta}}$ denote the MREL estimate based on the original data-set, and $l_j^*(\hat{\boldsymbol{\beta}})$ $(j = 1, 2, \ldots, 2000)$ be the 2000 bootstrap MREL ratio statistics. We shall take the 1900th statistic $l_{(1900)}^*(\hat{\boldsymbol{\beta}})$ as the 95 % quantile of the MREL method.

All confidence regions/intervals are re-computed, and presented in Fig. 3 and Table 5. It is clear from Fig. 3 that the MREL confidence region for (β_0, β_1) is significantly smaller than those based on the LS and LSEL. When only one component of (β_0, β_1) is of interest, we find from Table 5 that all the MREL confidence intervals are much shorter than those based on the LS and LSEL. These observations provide strong evidence for the priority of the MREL.

6 Concluding remarks

In this paper, in order to make inference about the regression coefficient of a linear regression model, we first investigate the properties of the modal regression with a fixed

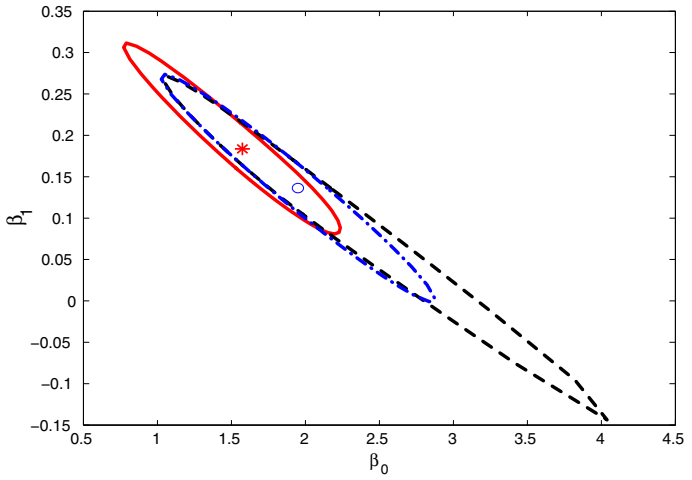

Fig. 3 Confidence regions based on 2000 bootstrap sampling for three different methods: LS (*blue dot dash line*); LSEL (*black dash line*); MREL (*red solid line*), where *circle* and *asterisk* denote the point estimate of LSE and MRE, respectively. (Color figure online)

Table 5 The confidence interval based on 2000 bootstrap resampling

Method	LS	LSEL	MREL
β_0	(1.1859, 2.7136)	(1.2267, 3.4541)	(1.0896, 2.0651)
β_1	(0.0205, 0.2520)	(−0.0614, 0.2421)	(0.1108, 0.2608)

bandwidth, then propose an empirical likelihood estimation approach based on modal regression estimation equation. It has been shown that the proposed estimator is more robust and efficient than the least square based methods for many non-normal error distributions or data containing outliers. Though our current research is focusing on linear regression, the framework can be extended to nonparametric or semi-parametric models, such as single-index models, partially linear models and semi-varying coefficient models. In addition, with high-dimensional covariates in regression models, sparse modeling is often considered superior, it is also interesting to consider robust penalized empirical likelihood based on modal regression, which can be taken as a future research topic.

Acknowledgments The research was supported in part by National Natural Science Foundation of China (11171112, 11001083, 11371142), Chinese Ministry of Education the 111 Project (B14019), Doctoral Fund of Ministry of Education of China (20130076110004), The Natural Science Project of Jiangsu Province Education Department (13KJB110024) and Natural Science Fund of Nantong University (13ZY001).

Appendix

Proof of Theorem 1

Proof We first prove the root-n consistency of $\hat{\boldsymbol{\beta}}$, i.e., $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$. It is sufficient to show that for any given $\varrho > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\boldsymbol{v}\|=C} Q_h(\boldsymbol{\beta}_0+n^{-1/2}\boldsymbol{v}) < Q_h(\boldsymbol{\beta}_0)\right\} \geq 1-\varrho, \tag{14}$$

where the function $Q_h(\cdot)$ is defined in (2).

For any vector $\boldsymbol{v}$ with length C, by the second-order Taylor expansion, we have

$$\begin{aligned}
&nQ_h\left(\boldsymbol{\beta}_0+n^{-1/2}\boldsymbol{v}\right)-nQ_h(\boldsymbol{\beta}_0)\\
&=\sum_{i=1}^{n}\left\{\phi_h\left(\epsilon_i-n^{-1/2}\boldsymbol{x}_i^T\boldsymbol{v}\right)-\phi_h(\epsilon_i)\right\}\\
&=-\sum_{i=1}^{n}\phi_h'(\epsilon_i)n^{-1/2}\boldsymbol{x}_i^T\boldsymbol{v}+\sum_{i=1}^{n}\frac{1}{2}\phi_h''(\epsilon_i)\left(n^{-1/2}\boldsymbol{x}_i^T\boldsymbol{v}\right)^2\\
&\quad-\sum_{i=1}^{n}\frac{1}{6}\phi_h'''(\xi_i)\left(n^{-1/2}\boldsymbol{x}_i^T\boldsymbol{v}\right)^3\\
&\equiv I_1+I_2+I_3,
\end{aligned} \tag{15}$$

where ξ_i lies between ϵ_i and $\epsilon_i-n^{-1/2}\boldsymbol{x}_i^T\boldsymbol{v}$.

We study respectively the magnitudes of I_1, I_2 and I_3. Let $A_n=\sum_{i=1}^{n}\phi_h'(\epsilon_i)n^{-1/2}\boldsymbol{x}_i$. It follows from condition (C1) and $\mathrm{E}(\phi_h'(\epsilon))=0$ that,

$$\mathrm{Var}\,(A_n)=\mathrm{E}\left\{\phi_h'(\epsilon_i)\right\}^2\mathrm{Var}(\boldsymbol{x}_i)=G(h)\Sigma. \tag{16}$$

The finiteness of $\mathrm{Var}(\boldsymbol{x}_i)$ and $G(h)=\boldsymbol{E}(\phi'(\epsilon)^2)$ implies that

$$\max_{1\leq i\leq n}\left|\phi_h'(\epsilon_i)n^{-1/2}\boldsymbol{x}_i\right|=o_p(1). \tag{17}$$

Then by central limit theorem, we have for fixed C that $A_n\xrightarrow{d}N(0,G(h)\Sigma)$, and therefore $I_1\xrightarrow{d}N(0,G(h)\boldsymbol{v}^T\Sigma\boldsymbol{v})$.

For I_2, with the strong law of large numbers, we have $I_2=\frac{1}{2}F(h)\boldsymbol{v}^T\Sigma\boldsymbol{v}+o(1)$, where $F(h)$ is defined in condition (C1).

About I_3, we find that

$$\begin{aligned}
|I_3|&\leq\left|\sum_{i=1}^{n}\frac{1}{6}\phi_h'''(\xi_i)(n^{-1/2}\boldsymbol{x}_i^T\boldsymbol{v})^2\right|\cdot\max_{1\leq i\leq n}\left(|\boldsymbol{x}_i^T\boldsymbol{v}|/\sqrt{n}\right)\\
&\leq\left|\frac{1}{6n}\sum_{i=1}^{n}\rho_{h,c}(\epsilon_i)(\boldsymbol{x}_i^T\boldsymbol{v})^2\right|\cdot\max_{1\leq i\leq n}\left(\|\boldsymbol{x}_i\|/\sqrt{n}\right)\|\boldsymbol{v}\|.
\end{aligned} \tag{18}$$

Condition (C2) implies that $\frac{1}{6n}\sum_{i=1}^{n}\rho_{h,c}(\epsilon_i)(\boldsymbol{x}_i^T\boldsymbol{v})^2 = O_p(1)$. It then follows from the fact that $\max_{1\le i\le n}(\|\boldsymbol{x}_i\|/\sqrt{n}) = o_p(1)$ that

$$I_3 = O_p(1)\cdot o_p(1)\cdot O_p(1) = o_p(1).$$

Overall, we obtain that for any $\boldsymbol{v}$ with $\|\boldsymbol{v}\| = C$,

$$nQ_h(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{v}) - nQ_h(\boldsymbol{\beta}_0) = -A_n^T\boldsymbol{v} + (1/2)F(h)\boldsymbol{v}^T\Sigma\boldsymbol{v} + \delta_n$$

with $\delta_n = o_p(1)$. The fact $-A_n^T\boldsymbol{v} \xrightarrow{d} N(0, G(h)\boldsymbol{v}^T\Sigma\boldsymbol{v})$ implies that for any $\varrho > 0$ and any nonzero $\boldsymbol{v}$, there exists $K > 0$ such that

$$P\left(\left|A_n^T\boldsymbol{v}\right| < K\sqrt{G(h)\boldsymbol{v}^T\Sigma\boldsymbol{v}}\right) > 1-\varrho.$$

Thus with probability $1-\varrho$, it holds that

$$nQ_h\left(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{v}\right) - nQ_h(\boldsymbol{\beta}_0) \le K\sqrt{G(h)\boldsymbol{v}^T\Sigma\boldsymbol{v}} + (1/2)F(h)\boldsymbol{v}^T\Sigma\boldsymbol{v} + \delta_n.$$

Note that $F(h) < 0$. Clearly, when n and C are both large enough,

$$K\sqrt{G(h)\boldsymbol{v}^T\Sigma\boldsymbol{v}} + (1/2)F(h)\boldsymbol{v}^T\Sigma\boldsymbol{v} + \delta_n < 0.$$

In summary, for any $\varrho > 0$, there exists $C > 0$ such that for $\boldsymbol{v} = C$, $nQ_h(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{v}) - nQ_h(\boldsymbol{\beta}_0)$ is negative with probability at least $1-\varrho$. Thus, (14) holds. That is, with the probability approaching 1, there exists a local maximizer $hat\boldsymbol{\beta}$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(1/\sqrt{n})$.

We turn to proving the asymptotical normality of $\hat{\boldsymbol{\beta}}$. Denote $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, then $\hat{\boldsymbol{\gamma}}$ satisfies the following equation

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\phi_h'(\epsilon_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\gamma}}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\left\{\phi_h'(\epsilon_i) - \phi_h''(\epsilon_i)\boldsymbol{x}_i^T\hat{\boldsymbol{\gamma}} + \frac{1}{2}\phi_h'''(\epsilon_i^*)\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\gamma}}\right)^2\right\} \\ &\triangleq J_1 + J_2\hat{\gamma} + J_3, \end{aligned} \tag{19}$$

where ϵ_i^* lies between ϵ_i and $\epsilon_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\gamma}}$. We have shown that

$$\sqrt{n}J_1 \xrightarrow{\mathrm{d}} N(0, G(h)\boldsymbol{\Sigma}), \quad J_2 \xrightarrow{\mathrm{p}} F(h)\boldsymbol{\Sigma}.$$

Meanwhile the fact $\hat{\boldsymbol{\gamma}} = O_p(n^{-1/2})$ and condition (C2) implies that $J_3 = o_p(1)$. Thus Eq. (19) implies $\hat{\boldsymbol{\gamma}} = -J_2^{-1}J_1 + o_p(1)$. Since the bandwidth h is a constant not

depending on n, by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \sqrt{n}\hat{\boldsymbol{\gamma}} \xrightarrow{\mathrm{d}} N(0, \boldsymbol{\Sigma}^{-1})\{G(h)/F^2(h)\}.$$

□

The following lemma is needed to prove Theorem 2.

Lemma 1 *Under the conditions of Theorem 1, the λ_{β_0} in (10) satisfies $\|\boldsymbol{\lambda}_{\beta_0}\| = O_p(n^{-1/2})$.*

Proof Denote $\boldsymbol{\lambda}_{\beta_0} = \zeta \mathbf{u}_0$ with $\mathbf{u}_0$ a unit vector and $\zeta = \|\boldsymbol{\lambda}_{\beta_0}\|$. Define matrix $\boldsymbol{\Phi}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \xi_i(\boldsymbol{\beta})\xi_i^T(\boldsymbol{\beta})$ and $Z = \max_{1\le i\le n}\|\xi_i(\boldsymbol{\beta}_0)\|$. It follows from the definition of $\boldsymbol{\lambda}_{\beta_0}$ that

$$\begin{aligned} 0 &= \frac{\mathbf{u}_0^T}{n}\sum_{i=1}^n \frac{\xi_i(\boldsymbol{\beta}_0)}{1+\zeta\mathbf{u}_0^T\xi_i(\boldsymbol{\beta}_0)} = \frac{\mathbf{u}_0^T}{n}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0) - \frac{\zeta}{n}\sum_{i=1}^n \frac{\{\mathbf{u}_0^T\xi_i(\boldsymbol{\beta}_0)\}^2}{1+\zeta\mathbf{u}_0^T\xi_i(\boldsymbol{\beta}_0)} \\ &\le \frac{\mathbf{u}_0^T}{n}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0) - \frac{\zeta}{1+\zeta Z}\frac{1}{n}\sum_{i=1}^n (\mathbf{u}_0^T\xi_i(\boldsymbol{\beta}_0))^2 \\ &= \frac{\mathbf{u}_0^T}{n}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0) - \frac{\zeta}{1+\zeta Z}\mathbf{u}_0^T\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)\mathbf{u}_0, \end{aligned}$$

which implies

$$\zeta\left\{\mathbf{u}_0^T\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)\mathbf{u}_0 - Z\frac{\mathbf{u}_0^T}{n}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0)\right\} \le \frac{\mathbf{u}_0^T}{n}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0). \tag{20}$$

By the Cauchy–Schwarz inequality and law of large numbers, we have

$$\left|\frac{\mathbf{u}_0^T}{n}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0)\right| \le \left\|\frac{1}{n}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0)\right\| = O_p(n^{-1/2}). \tag{21}$$

This together with Eq. (17) gives

$$Z\frac{\mathbf{u}_0^T}{n}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0) = o_p(1). \tag{22}$$

Condition (C1) and law of large numbers implies $\boldsymbol{\Phi}_n \xrightarrow{\mathrm{p}} G(h)\boldsymbol{\Sigma}$, which means that there exists $c > 0$ such that $P(\boldsymbol{u}_0^T\boldsymbol{\Phi}_n\boldsymbol{u}_0 > c) \to 1$ as $n \to \infty$.

Furthermore, since $n^{-1/2}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0) \xrightarrow{\mathrm{d}} N(0, \boldsymbol{\Phi})$, we find that $\|\boldsymbol{\lambda}_{\beta_0}\| = O_p(n^{-1/2})$. □

Proof of Theorem 2

Proof Let $y_i = \boldsymbol{\lambda}_{\beta_0}^T \xi_i(\boldsymbol{\beta}_0)$. It follows from Lemma 1 that

$$\max_{1\le i\le n} |y_i| \le \|\boldsymbol{\lambda}_{\beta_0}\| \max_{1\le i\le n} |\xi_i(\boldsymbol{\beta}_0)| = O_p(n^{-1/2}) o_p(n^{1/2}) = o_p(1),$$

which implies that the upcoming Taylor expansion is valid. Applying the second-order Taylor expansion on $(1+y_i)^{-1}$ for i from 1 to n, we obtain from Eq. (10) that

$$\boldsymbol{\lambda}_{\beta_0} = \{\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)\}^{-1} \frac{1}{n} \sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0) + \{\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)\}^{-1} r_n(\boldsymbol{\beta}_0), \tag{23}$$

where $r_n(\boldsymbol{\beta}_0) = (1/n)\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0)(1+\delta_i^*)^{-1}\{\boldsymbol{\lambda}_{\beta_0}^T \xi_i(\boldsymbol{\beta}_0)\}^2$ and δ_i^* lies between 0 and y_i. Clearly $\max_{1\le i\le n} |\delta_i^*| = o_p(1)$. Therefore

$$\begin{aligned} |r_n(\boldsymbol{\beta}_0)| &\le \max_{1\le i\le n} \|\xi_i(\boldsymbol{\beta}_0)\| (1 - \max_{1\le i\le n} |\delta_i^*|)^{-1} \boldsymbol{\lambda}_{\beta_0}^T \boldsymbol{\Phi}_n(\boldsymbol{\beta}_0) \boldsymbol{\lambda}_{\beta_0} \\ &= o_p(n^{1/2}) O_p(n^{-1}) = o_p(n^{-1/2}). \end{aligned}$$

Thus we have

$$\boldsymbol{\lambda}_{\beta_0} = \{\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)\}^{-1} \frac{1}{n} \sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0) + o_p(n^{-1/2}). \tag{24}$$

Similarly, by the third-order Taylor expansion on $\log(1+y_i)$ for all i, we have

$$\begin{aligned} -2l(\boldsymbol{\beta}_0) = 2\sum_{i=1}^n \boldsymbol{\lambda}_{\beta_0}^T \xi_i(\boldsymbol{\beta}_0) - \sum_{i=1}^n \left\{\boldsymbol{\lambda}_{\beta_0}^T \xi_i(\boldsymbol{\beta}_0)\right\}^2 \\ + \tfrac{2}{3}\sum_{i=1}^n \left\{\boldsymbol{\lambda}_{\beta_0}^T \xi_i(\boldsymbol{\beta}_0)\right\}^3 (1+\eta_i^*)^{-3} \end{aligned} \tag{25}$$

where η_i lies between 0 and y_i. It can be verified that

$$\begin{aligned} &\left| \sum_{i=1}^n \left\{\boldsymbol{\lambda}_{\beta_0}^T \xi_i(\boldsymbol{\beta}_0)\right\}^3 (1+\eta_i^*)^{-3} \right| \\ &\le \max_{1\le i\le n} \left|\boldsymbol{\lambda}_{\beta_0}^T \xi_i(\boldsymbol{\beta}_0)\right| \left(1 - \max_{1\le i\le n} |\eta_i^*|\right)^{-3} n \boldsymbol{\lambda}_{\beta_0}^T \boldsymbol{\Phi}_n(\boldsymbol{\beta}_0) \boldsymbol{\lambda}_{\beta_0} \\ &= o_p(1) \cdot O_p(1) = o_p(1). \end{aligned}$$

Furthermore, by incorporating Eq. (24), we have

$$-2l(\boldsymbol{\beta}_0) = \left\{n^{-1/2}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0)\right\}^T \left\{\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)\right\}^{-1} \left\{n^{-1/2}\sum_{i=1}^n \xi_i(\boldsymbol{\beta}_0)\right\} + o_p(1). \tag{26}$$

Since $\xi_i(\boldsymbol{\beta}_0) = \boldsymbol{x}_i \phi_h' \epsilon_i$, it follows from conclusion of Lemma 1 that as $n \to \infty$,

$$\left\{ n^{-1/2} \sum_{i=1}^{n} \xi_i(\boldsymbol{\beta}_0) \right\}^T \left\{ \boldsymbol{\Phi}_n(\boldsymbol{\beta}_0) \right\}^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \xi_i(\boldsymbol{\beta}_0) \right\} \xrightarrow{\mathrm{d}} \chi_p^2,$$

which immediately implies $-2l(\boldsymbol{\beta}_0) \xrightarrow{\mathrm{d}} \chi_p^2$. This completes the proof. □

References

Chatterjee S, Price B (1977) Regression analysis by example. Wiley, New York

Chen J, Variyath AM, Abraham B (2008) Adjusted empirical likelihood and its properties. J. Comput. Graph. Stat. 17:426–443

Chen S, Keilegom I (2009) A review on empirical likelihood methods for regression (with discussions). Test 18:415–447

Chen X, Wang Z, Martin J (2010) Asymptotic analysis of robust lassos in the presence of noise with large variance. IEEE Trans. Inf. Theory 56:5131–5149

Huber P (1981) Robust Statistic. Wiley, New York

Johnson B, Peng L (2008) Rank-based variable selection. J. Nonparametr. Stat. 20:241–252

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46:33–50

Lee M (1989) Mode regression. J. Econom. 42:337–349

Liu Y, Chen J (2010) Adjusted empirical likelihood with high-order precision. Ann. Stat. 38:1341–1362

Parzen E (1962) On estimation of a probability density function and mode. Ann. Math. Stat. 33:1065–1076

Owen A (1988) Empirical likelihood ratio confidence intervals for a single function. Biometrika 75:237–249

Owen A (1990) Empirical likelihood ratio confidence regions. Ann. Stat. 18:90–120

Owen A (1991) Empirical likelihood for linear models. Ann. Stat. 19:1725–1747

Owen A (2001) Empirical Likelihood. Chapman and Hall, New York

Qin J, Lawless J (1994) Empirical likelihood and general estimating equations. Ann. Stat. 22:300–325

Rousseeuw P, Leroy A (1987) Robust Regression and Outlier Detection. Wiley, New York

Scott D (1992) Multivariate Density Estimation: Theory, Practice and Visualization. Wiley, New York

Wei C, Luo Y, Wu X (2012) Empirical likelihood for partially linear additive errors-in-variables models. Stat. Pap. 53:48–496

Yao, W., Li, L.: A new regression model: modal linear regression. Scand. J. Stat. (2013). doi:10.1111/sjos.12054

Yao W, Lindsay B, Li R (2012) Local modal regression. J. Nonparametr. Stat. 24:647–663

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

Zi X, Zou C, Liu Y (2012) Two-sample empirical likelihood method for difference between coefficients in linear regression model. Stat. Pap. 53:83–93

Zou H, Yuan M (2008) Composite quantile regression and the oracle model selection theory. Ann. Stat. 36:1108–1126