Journal of Applied Mathematics and Physics, 2014, 2, 948-952
Published Online September 2014 in SciRes. http://www.scirp.org/journal/jamp
http://dx.doi.org/10.4236/jamp.2014.210107

Scientific Research

Empirical Likelihood Diagnosis of Modal Linear Regression Models

Shuling Wang[1], Lin Zheng[2], Jiangtao Dai[3]

[1]Department of Fundamental Course, Air Force Logistics College, Xuzhou, China
[2]School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China
[3]Fundamental Science Department, North China Institute of Astronautic Engineering, Langfang, China
Email: 155328313@qq.com

Received 10 August 2014; revised 10 September 2014; accepted 17 September 2014

Abstract

In this paper, we investigate the empirical likelihood diagnosis of modal linear regression models. The empirical likelihood ratio function based on modal regression estimation method for the regression coefficient is introduced. First, the estimation equation based on empirical likelihood method is established. Then, some diagnostic statistics are proposed. At last, we also examine the performance of proposed method for finite sample sizes through simulation study.

Keywords

Modal Linear Regression Model, Empirical Likelihood, Outliers, Influence Analysis

1. Introduction

The mode of a distribution is regarded as an important feature of data. Several authors have made efforts to identify the modes of population distributions for low-dimensional data. See, for example, Muller and Sawitzki [1]; Scott [2]; Friedman and Fisher [3]; Chaudhuri and Marron [4]; Fisher and Marron [5]; Davies and Kovac [6] Hall, Minnotte and Zhang [7]; Ray and Lindsay [8]; Yao and Lindsay [9]. In high-dimensional data, it is common to impose some model structure assumptions such as assumption on conditional distributions. Thus, it is of great interest to study the mode hunting for conditional distributions.

Given a random sample $\{(x_i, y_i), i = 1, \cdots, n\}$, where x_i is a p-dimension column vector, $f(y \mid x)$ is the conditional density function. For the conventional regression models, the mean of $f(y \mid x)$ is usually used to investigate the relationship between Y and x and the linear regression assumes that the mean of $f(y \mid x)$ is a linear function of x. Yao and Li [10] proposed a new regression model called modal linear regression that assumes the mode of $f(y \mid x)$ is a linear function of the predictor x. Modal linear regression measures the

How to cite this paper: Wang, S.L., Zheng, L. and Dai, J.T. (2014) Empirical Likelihood Diagnosis of Modal Linear Regression Models. *Journal of Applied Mathematics and Physics*, **2**, 948-952. http://dx.doi.org/10.4236/jamp.2014.210107

center using the "most likely" conditional values rather than the conditional average used by the traditional linear regression.

Lee [11] used the uniform kernel and Epanechnikov kernel to estimate the modal regression. However, their estimators are of little practical use because the object function is non-differentiable and its distribution is intractable. Scott [2] mentioned the modal regression, but little methodology is given on how to implement it in practice. Recently, Yao *et al*. [12] 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. Yao and Li [10] suggested using the Gaussian kernel and developed MEM algorithm to compute modal estimators for linear models. Their estimation procedure is very convenient to be implemented for practitioners and the result is encouraging for many non-normal error distributions. Yu and Aristodemou [13] studied modal regression from Bayesian perspective. In addition, Zhao, Zhang and Liu [14] considered how to yield a robust empirical likelihood estimation for regression models.

The empirical likelihood method originates from Thomas & Grunkemeier [15]. Owen [16] first proposed the definition of empirical likelihood and expounded the system info of empirical likelihood. Zhu and Ibrahim [17] utilized this method for statistical diagnostic, and they developed diagnostic measures for assessing the influence of individual observations when using empirical likelihood with general estimating equations, and used these measures to construct goodness-of-fit statistics for testing possible misspecification in the estimating equations. Liugen Xue and Lixing Zhu [18] summarized the application of this method.

Over the last several decades, the diagnosis and influence analysis of linear regression model has been fully developed (R. D. Cook and S. Weisberg [19], Bo-cheng Wei, Go-bin Lu & Jian-qing Shi [20]). So far the statistical diagnostics of modal linear regression models based on empirical likelihood method has not yet been seen in the literature. This paper attempts to study it.

The rest of the paper is organized as follows. In Section 2, we review the modal regression. In Section 3, empirical likelihood and estimation equation are presented. The main results are given in Section 4. Simulation study is given to illustrate our results in Section 5.

2. Modal Linear Regression

Suppose a response variable y given a set of predictor x is distributed with a probability density function (PDF) $f(y|x)$. Yao and Li [10] proposed to use the mode of $f(y|x)$, denoted by $\text{Mode}(Y|x) = \arg\max_y (f(y|x))$, to investigate the relationship between Y and x. The proposed modal linear regression method assumes that

$$\text{Mode}(Y|x) = x^{\mathrm{T}}\beta. \tag{1}$$

The idea of modal linear regression can be easily generalized to other models such as nonlinear regression, nonparametric regression, and varying coefficient partially linear regression. To include the intercept term in (1), we assume that the first element of x is 1. Let $\varepsilon = y - x^T\beta$ and denote by $g(\varepsilon|x)$ the conditional density of ε given x. Here, we allow the conditional density of ε given x to depend on x. Based on the model assumption (1), one knows that $g(\varepsilon|x)$ is maximized at 0 for any x. If $g(\varepsilon|x)$ is symmetric about 0, the β in (1) will be the same as the conventional linear regression parameters. However, if $g(\varepsilon|x)$ is skewed, they will be different and it is even possible that the modal regression is a linear function of x but the conventional mean regression function is nonlinear.

Yao and Li [10] proposed to estimate the modal regression parameter β in (1) by maximizing

$$Q_h(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n}\varphi_h(y_i - x_i^{\mathrm{T}}\beta) \tag{2}$$

where $\varphi_h(t) = h^{-1}\varphi(t/h)$ and $\varphi(t)$ is a kernel density function. Denote by $\hat{\beta}$ the maximizer of (2). We call $\hat{\beta}$ the modal linear regression (MODLR) estimator.

3. Empirical Likelihood and Estimation Equation

In this section, we review empirical likelihood based on modal regression for regression coefficients, then establish the estimation equations.

Similarly to Zhao, Zhang and Liu [14], we define an auxiliary random vector

$$\xi_i(\beta) = x_i \varphi_h'\left(y_i - x_i^{\mathrm{T}}\beta\right),\ i = 1, \cdots, n. \tag{3}$$

Note that $E\{\xi_i(\beta_0)\} = 0$, where β_0 is the true parameter value. According to the empirical likelihood principle, we define the empirical likelihood ratio function of β to be

$$l(\beta) = \sup\left\{\prod_{i=1}^{n}(np_i) \,\middle|\, p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \xi_i(\beta) = 0\right\}. \tag{4}$$

By the method of Lagrange multipliers, similar to that used in Owen (2001), $l(\beta)$ is well-defined and can be re-expressed as

$$l(\beta) = \prod_{i=1}^{n}\left\{1 + \lambda^T \xi_i(\beta)\right\}^{-1}, \tag{5}$$

where λ is determined by the constraint equation

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

Motivated by Zhu and Ibrahim [17], we regard λ and β as independent variables and define

$$Q_n(\lambda, \beta) = -n^{-1}\sum_{i=1}^{n}\log\left(1 + \lambda^T \xi_i(\beta)\right),$$

Obviously, the maximum empirical likelihood estimates $\hat{\beta}$ and $\hat{\lambda}$ are the solutions of following equations:

$$\begin{cases} Q_{1,n}(\lambda, \beta) = \dfrac{\partial Q_n(\lambda, \beta)}{\partial \lambda} = -n^{-1}\displaystyle\sum_{i=1}^{n}\xi_i(\beta)\left\{1 + \lambda^T \xi_i(\beta)\right\}^{-1} = 0 \\ Q_{2,n}(\lambda, \beta) = \dfrac{\partial Q_n(\lambda, \beta)}{\partial \beta} = -n^{-1}\displaystyle\sum_{i=1}^{n}\dfrac{\partial \xi_i(\beta)}{\partial \beta}\lambda\left\{1 + \lambda^T \xi_i(\beta)\right\}^{-1} = 0. \end{cases}$$

4. Local Influence Analysis of Model

We consider the local influence method for a case-weight perturbation $\omega \in R^n$, for which the empirical log-likelihood function $l_E(\xi \mid \omega)$ is defined by $l_E(\xi \mid \omega) = \sum_{i=1}^{n}\omega_i l_{E,i}(\xi)$. In this case, $\omega = \omega^0$, defined to be an $n \times 1$ vector with all elements equal to 1, represents no perturbation to the empirical likelihood, because $l_E(\xi \mid \omega^0) = l_E(\xi)$. Thus, the empirical likelihood displacement is defined as $l_{DE}(\omega) = 2\left[l_E\left(\hat{\xi}\right) - l_E\left\{\hat{\xi}(\omega)\right\}\right]$, where $\hat{\xi}(\omega)$ is the maximum empirical likelihood estimator of ξ based on $l_E(\xi \mid \omega)$. Let $\omega(a) = \omega^0 + ah$ with $\omega(0) = \omega^0$ and $\mathrm{d}\omega(a)/\mathrm{d}a\big|_{a=0} = h$, where h is a direction in R^n. Thus, the normal curvature of the influence graph $\left(\omega^{\mathrm{T}}, LD_E(\omega)\right)^{\mathrm{T}}$ is given by $\mathrm{C}_h(\omega^0) = h^{\mathrm{T}} H_{LD_E(\omega^0)} h$, where

$$H_{LD_E(\omega^0)} = -2\frac{\partial^2 LD_E\left\{\hat{\xi}(\omega)\right\}}{\partial\omega\partial\omega^T}\bigg|_{\omega^0} = 2\Delta^T\left\{-\partial_{\xi}^2 l_E(\xi)\right\}^{-1}\Delta\Big|_{\omega^0, \hat{\xi}}$$, in which $\Delta = \partial_{\xi\omega}^2 LD_E(\xi, \omega)$ is a $p \times n$ matrix with (k, i)-th element given by $\partial_{\xi_k} l_{E,i}(\xi)$.

We consider two local influence measures based on the normal curvature $\mathrm{C}_h(\omega^0)$ as follows. Let $\lambda_1 \geq \cdots \geq \lambda_p \geq \lambda_{p+1} = \cdots = \lambda_n = 0$ be the ordered eigenvalues of the matrix $H_{LD_E(\omega^0)}$ and let $\left\{v_m = (v_{m1}, \cdots, v_{mn})^{\mathrm{T}} : m = 1, \cdots n\right\}$ be the associated orthonormal basis, that is, $H_{LD_E(\omega^0)} v_m = \lambda_m v_m$. Thus, the spectral decomposition of $H_{LD_E(\omega^0)}$ is given by

$$H_{LD_E(\omega^0)} = \sum_{m=1}^{n}\lambda_m v_m v_m^{\mathrm{T}}.$$

The most popular local influence measures include v_1, which corresponds the largest eigen value λ_1, as well

as $C_{e_j} = \sum_{m=1}^{p} \lambda_m v_{mj}^2$, where e_j is an $n \times 1$ vector with j-th component 1 and 0 otherwise. The v_1 represents the most influential perturbation to the empirical likelihood function, whereas the j-th observation with a large C_{e_j} can be regarded as influential.

As the discuss of Zhu *et al*. [17], for varying-coefficient density-ratio model, we can deduce that

$$C_{e_j} = 2ELD_j\left\{1+o_p(1)\right\} = 2ECD_j\left\{1+o_p(1)\right\} = -2n^{-1}\Delta_j^T S_{22.1}^{-1}\Delta_j\left\{1+o_p(1)\right\}, \tag{4}$$

where $\Delta_j = \partial_\beta l_{E,j}(\beta)\big|_{\beta=\hat{\beta}} = \dfrac{S_{21}S_{11}^{-1}\xi_j(\beta)}{1+\hat{\lambda}^{\mathrm{T}}\xi_j(\beta)} + o_p(1)$, $S_{11} = \partial_\lambda Q_{1,n} = \dfrac{1}{n}\sum_{i=1}^{n}\dfrac{\xi_i(\beta)\xi_i(\beta)^T}{\left(1+\lambda^T\xi_i(\beta)\right)^2}\Big|_{\beta=\hat{\beta},\lambda=\hat{\lambda}}$,

$$S_{12} = \partial_\beta Q_{1,n} = \frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i(\beta)\lambda^{\mathrm{T}}\partial_\beta\left(\xi_i(\beta)\right) - \partial_\beta\left(\xi_i(\beta)\right)\left(1+\lambda^T\xi_i(\beta)\right)}{\left(1+\lambda^T\xi_i(\beta)\right)^2}\Bigg|_{\beta=\hat{\beta},\lambda=\hat{\lambda}},$$

$$S_{21} = \partial_\lambda Q_{2,n} = \frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i(\beta)\lambda^{T}\partial_\beta\left(\xi_i(\beta)\right) - \partial_\beta\left(\xi_i(\beta)\right)\left(1+\lambda^T\xi_i(\beta)\right)}{\left(1+\lambda^T\xi_i(\beta)\right)^2}\Bigg|_{\beta=\hat{\beta},\lambda=\hat{\lambda}},$$

$$S_{22} = \partial_\beta Q_{2,n} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial_\beta^T\left(\xi_i(\beta)\right)\lambda\lambda^T\partial_\beta\left(\xi_i(\beta)\right)}{\left(1+\lambda^T\xi_i(\beta)\right)^2}\Bigg|_{\beta=\hat{\beta},\lambda=\hat{\lambda}}, \quad S_{22.1} = -S_{21}S_{11}^{-1}S_{12}.$$

5. Numerical Study

We generate data-sets from following model

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \varepsilon_i,\ i = 1,\cdots,n$$

where the covariates $x_i = \left(x_{i1}, x_{i2}, x_{i3}\right)^{\mathrm{T}}$ 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 $\beta = \left(\beta_0,\cdots,\beta_3\right)^T = (1.5, 2, -1.2, 0)^T$. The error ε_i is independent of x_i. For ease of computation, we use the standard normal density function for $\varphi(t)$. Simulation results are computed based on 1000 random samples with the sample size being 150.

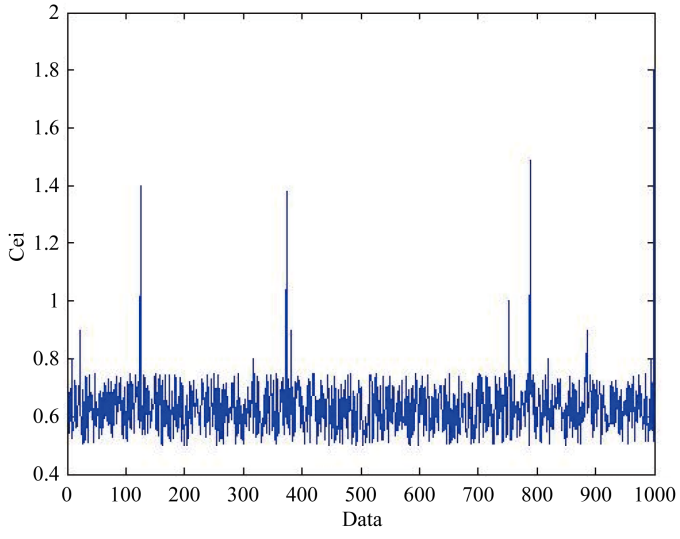

Figure 1. The influence value of C_{e_i}.

In order to check out the validity of our proposed methodology, we change the value of the first, 125th, 374th, 789th and 999th data. For every case, it is easy to obtain $\xi_i(\beta)$. For β and λ, using the samples, we evaluated their maximum empirical likelihood estimators.

Consequently, it is easy to calculate the value of $S_{11}, S_{12}, S_{21}, S_{22}$ and C_{e_i}. The result of C_{e_i} is as **Figure 1**.

From the figure, we can see that in most cases, the value of C_{e_i} are reasonably close to one fixed value. Following the definition and properties of C_{e_i}, we can diagnose the strong influence points, the value of which deviate from the average seriously. It can be seen from the result of C_{e_i} that the first, 125th, 374th, 789th and 999th data are strong influence points. Indeed, our results are illustrated.

6. Discussion

In this paper, we considered the statistical diagnosis for modal linear regression models based on empirical likelihood. Through simulation study, we illustrate that our proposed method can work fairly well.

References

[1] Muller, D.W. and Sawitzki, G. (1991) Excess Mass Estimates and Tests for Multimodality. *Journal of the American Statistical Association*, **86**, 738-746.

[2] Scott, D.W. (1992) Multivariate Density Estimation: Theory, Practice and Visualization. Wiley, New York. http://dx.doi.org/10.1002/9780470316849

[3] Friedman, J.H. and Fisher, N.I. (1999) Bump Hunting in High-Dimensional Data. *Statistics and Computing*, **9**, 123-143. http://dx.doi.org/10.1023/A:1008894516817

[4] Chaudhuri, P. and Marron, J.S. (1999) Sizer for Exploration of Structures in Curves. *Journal of the American Statistical Association*, **94**, 807-823. http://dx.doi.org/10.1080/01621459.1999.10474186

[5] Fisher, N.I. and Marron, J.S. (2001) Mode Testing via the Excess Mass Estimate. *Biometrika*, **88**, 499-517. http://dx.doi.org/10.1093/biomet/88.2.499

[6] Davies, P.L. and Kovac, A. (2004) Densities, Spectral Densities and Modality. *Annals of Statistics*, **32**, 1093-1136.

[7] Hall, P., Minnotte, M.C. and Zhang, C. (2004) Bump Hunting with Non-Gaussian Kernels. *Annals of Statistics*, **32**, 2124-2141. http://dx.doi.org/10.1214/009053604000000715

[8] Ray, S. and Lindsay, B.G. (2005) The Topography of Multivariate Normal Mixtures. *Annals of Statistics*, **33**, 2042-2065. http://dx.doi.org/10.1214/009053605000000417

[9] Yao, W. and Lindsay, B.G. (2009) Bayesian Mixture Labeling by Highest Posterior Density. *Journal of American Statistical Association*, **104**, 758-767. http://dx.doi.org/10.1198/jasa.2009.0237

[10] Yao, W. and Li, L. (2013) A New Regression Model: Modal Linear Regression. *Scandinavian Journal of Statistics*, **41**, 656-671. http://dx.doi.org/10.1111/sjos.12054

[11] Lee, M.J. (1989) Mode Regression. *Journal of Econometrics*, **42**, 337-349. http://dx.doi.org/10.1016/0304-4076(89)90057-2

[12] Yao, W., Lindsay, B. and Li, R. (2012) Local Modal Regression. *Journal of Nonparametric Statistics*, **24**, 647-663. http://dx.doi.org/10.1080/10485252.2012.678848

[13] Yu, K. and Aristodemou, K. (2012) Bayesian Mode Regression. Technical Report. arXiv: 1208.0579v1.

[14] Zhao, W.H., Zhang, R.Q., Liu, Y.K. and Liu, J.C. (2014) Empirical Likelihood Based Modal Regression. Statistical Papers.

[15] Thomas, D.R. and Grunkemeier, G.L. (1975) Confidence Interval Estimation of Survival Interval Estimation of Survival Probabilities for Censored Data. *Journal of the American Statistical Association*, **70**, 865-871. http://dx.doi.org/10.1080/01621459.1975.10480315

[16] Owen, A. (2001) Empirical Likelihood. Chapman and Hall, New York. http://dx.doi.org/10.1201/9781420036152

[17] Zhu, H.T., Ibrahim, J.G., Tang, N.S. and Zhang, H.P. (2008) Diagnostic Measures for Empirical Likelihood of Generalized Estimating Equations. *Biometrika*, **95**, 489-507. http://dx.doi.org/10.1093/biomet/asm094

[18] Xue, L.G. and Zhu, L.X. (2010) Empirical Likelihood in Nonparametric and Semiparametric Models. Science Press, Beijing.

[19] Cook, R.D. and Weisberg, S. (1982) Residuals and Influence in Regression. Chapman and Hall, New York.

[20] Wei, B.C., Lu, G.B. and Shi, J.Q. (1990) Statistical Diagnostics. Publishing House of Southeast University, Nanjing.

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.