Test (2009) 18: 415–447
DOI 10.1007/s11749-009-0159-5

INVITED PAPER

# A review on empirical likelihood methods for regression

**Song Xi Chen · Ingrid Van Keilegom**

Received: 31 March 2009 / Accepted: 13 September 2009 / Published online: 4 November 2009
© Sociedad de Estadística e Investigación Operativa 2009

**Abstract** We provide a review on the empirical likelihood method for regression-type inference problems. The regression models considered in this review include parametric, semiparametric, and nonparametric models. Both missing data and censored data are accommodated.

**Keywords** Censored data · Empirical likelihood · Missing data · Nonparametric regression · Parametric regression · Semiparametric regression · Wilks' theorem

**Mathematics Subject Classification (2000)** 62-02 · 62E20 · 62F03 · 62G08 · 62G10 · 62J02 · 62N01

## 1 Introduction

It has been twenty years since Art Owen published his seminal paper (Owen 1988) that introduces the notion of empirical likelihood (EL). Since then, there has been a rich body of literature on the novel idea of formulating versions of nonparametric

This invited paper is discussed in the comments available at: doi:10.1007/s11749-009-0160-z, doi:10.1007/s11749-009-0161-y, doi:10.1007/s11749-009-0162-x, doi:10.1007/s11749-009-0163-9, doi:10.1007/s11749-009-0164-8, doi:10.1007/s11749-009-0165-7.

S.X. Chen
Department of Statistics, Iowa State University, Ames, IA 50011-1210, USA
e-mail: songchen@iastate.edu

S.X. Chen
Guanghua School of Management, Peking University, Beijing, China
e-mail: csx@gsm.pku.edu.cn

I. Van Keilegom (✉)
Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium
e-mail: ingrid.vankeilegom@uclouvain.be

likelihood in various settings of statistical inference. There have been two major reviews on the empirical likelihood. The first review was given by Hall and La Scala (1990) in the early years of the EL method, which summarized some key properties of the method. The second one was the book by the inventor of the methodology (Owen 2001), which provided a comprehensive overview up to that time.

The body of empirical likelihood literature is increasing rapidly, and it would be a daunting task to review the entire field in one review paper like this one. We therefore decided to concentrate our review on regression due to its prominence in statistical inference. The regression models considered in this review cover parametric, nonparametric, and semiparametric regression models. In addition to the case of completely observed data, we also accommodate missing and censored data in this review.

The EL method (Owen 1988, 1990) owns its broad usage and fast research development to a number of important advantages. Generally speaking, it combines the reliability of nonparametric methods with the effectiveness of the likelihood approach. It yields confidence regions that respect the boundaries of the support of the target parameter. The regions are invariant under transformations and behave often better than confidence regions based on asymptotic normality when the sample size is small. Moreover, they are of natural shape and orientation since the regions are obtained by contouring a log likelihood ratio, and they often do not require the estimation of the variance, as the studentization is carried out internally via the optimization procedure. The EL method turns out appealing not only in getting confidence regions, but it also has its unique attractions in parameter estimation and formulating goodness-of-fit tests.

## 2 Parametric regression

Suppose that we observe a sample of independent observations $\{(X_i^T, Y_i)^T\}_{i=1}^n$, where each $Y_i$ is regarded as the response of a $d$-dimensional design (covariate) variable $X_i$. The preliminary interest here is in the conditional mean function (regression function) of $Y_i$ given $X_i$. One distinguishes between the design $X_i$ being either fixed or random. Despite regression is conventionally associated with fixed designs, for ease of presentation, we will concentrate on random designs. The empirical likelihood analysis for fixed designs can be usually extended by regularizing the random designs.

Consider first the following parametric regression model:

$$Y_i = m(X_i; \beta) + \varepsilon_i \quad \text{for } i = 1, \ldots, n, \tag{1}$$

where $m(x; \beta)$ is the known regression function with an unknown $p$-dimensional ($p < n$) parameter $\beta \in R^p$, and the errors $\varepsilon_i$ are independent random variables such that $E(\varepsilon_i | X_i) = 0$ and $\mathrm{Var}(\varepsilon_i | X_i) = \sigma^2(X_i)$ for some function $\sigma(\cdot)$. Hence, the errors can be heteroscedastic. We require, like in all empirical likelihood formulations, that the errors $\epsilon_i$ have finite conditional variance, which is a minimum condition needed by the empirical likelihood method to ensure a limiting chi-square distribution for the empirical likelihood ratio.

The parametric regression function includes as special cases (i) the linear regression with $m(x;\beta) = x^T\beta$; (ii) the generalized linear model (McCullagh and Nelder, 1983) with $m(x;\beta) = G(x^T\beta)$ and $\sigma^2(x) = \sigma_0^2 V\{G(x^T\beta)\}$ for a known link function $G$, a known variance function $V(\cdot)$, and an unknown constant $\sigma_0^2 > 0$. Note that for these two special cases, $p = d$.

In the absence of model information on the conditional variance, the least squares (LS) regression estimator of $\beta$ is obtained by minimizing the sum of least squares

$$S_n(\beta) =: \sum_{i=1}^{n} \big\{Y_i - m(X_i;\beta)\big\}^2.$$

The LS estimator of $\beta$ is $\hat{\beta}_{ls} = \arg\inf_\beta S_n(\beta)$. When the regression function $m(x;\beta)$ is smooth enough with respect to $\beta$, $\hat{\beta}_{ls}$ will be a solution of the following estimating equation:

$$\sum_{i=1}^{n} \frac{\partial m(X_i;\beta)}{\partial \beta} \big\{Y_i - m(X_i;\beta)\big\} = 0. \tag{2}$$

Suppose that $\beta_0$ is the true parameter value such that it is the unique value to make $E[\frac{\partial m(X_i;\beta)}{\partial \beta}\{Y_i - m(X_i;\beta)\}|X_i] = 0$. Let $p_1, \ldots, p_n$ be a set of probability weights allocated to the data. The empirical likelihood (EL) for $\beta$, in the spirit of Owen (1988 1991), is

$$L_n(\beta) = \max \prod_{i=1}^{n} p_i, \tag{3}$$

where the maximization is subject to the constraints

$$\sum_{i=1}^{n} p_i = 1 \quad \text{and} \tag{4}$$

$$\sum_{i=1}^{n} p_i \frac{\partial m(X_i;\beta)}{\partial \beta} \big\{Y_i - m(X_i;\beta)\big\} = 0. \tag{5}$$

The empirical likelihood, as conveyed by (3), is essentially a constrained profile likelihood, with a trivial constraint (4) indicating the $p_i$'s are probability weights. The constraint (5) is the most important one as it defines the nature of the parameters. This formulation is similar to the original one given in Owen (1988, 1990) for the mean parameter, say $\mu$, of $X_i$. There the second constraint, reflecting the nature of $\mu$, was given by $\sum_{i=1}^{n} p_i(X_i - \mu) = 0$.

In getting the empirical likelihood at each candidate parameter value $\beta$, the above optimization problem as given in (3), (4), and (5) has to be solved for the optimal $p_i$'s. It may be surprising in first instance that the above optimization problem can admit a solution as there are $n$ $p_i$'s to be determined with only $p + 1$ constraints. As the objective function $L_n(\beta)$ is concave, and the constraints are linear in the $p_i$'s, the optimization problem does admit unique solutions.

The algorithm for computing $L_n(\beta)$ at a candidate $\beta$ is as follows. If the convex hull of the set of points (depending on $\beta$) $\{\frac{\partial m(X_i;\beta)}{\partial \beta}\{Y_i - m(X_i;\beta)\}\}_{i=1}^n$ in $R^p$ contains the origin (zero) in $R^p$, then the EL optimization problem for $L_n(\beta)$ admits a solution. If the zero of $R^p$ is not contained in the convex hull of the points for the given $\beta$, then $L_n(\beta)$ does not admit a finite solution as some weights $p_i$ are forced to take negative values; see Owen (1988, 1990) for a discussion on this aspect. Tsao (2004) studied the probability of the EL not admitting a finite solution and the dependence of this probability on dimensionality.

By introducing the Lagrange multipliers $\lambda_0 \in R$ and $\lambda_1 \in R^p$, the constrained optimization problem (3)–(5) can be translated into an unconstrained one with objective function

$$T(\mathbf{p}, \lambda_0, \lambda_1) = \sum_{i=1}^n \log(p_i) + \lambda_0 \left( \sum_{i=1}^n p_i - 1 \right) + \lambda_1^T \sum_{i=1}^n p_i \frac{\partial m(X_i;\beta)}{\partial \beta} \{Y_i - m(X_i;\beta)\}, \tag{6}$$

where $\mathbf{p} = (p_1, \ldots, p_n)^T$. Differentiating $T(\mathbf{p}, \lambda_0, \lambda_1)$ with respect to each $p_i$ and setting the derivative to zero, it can be shown after some algebra that $\lambda_0 = -n$, and by defining $\lambda = -n\lambda_1$, we find that the optimal $p_i$'s are given by

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta} \{Y_i - m(X_i;\beta)\}},$$

where, from the structural constraint (5), $\lambda$ satisfies

$$\sum_{i=1}^n \frac{\frac{\partial m(X_i;\beta)}{\partial \beta} \{Y_i - m(X_i;\beta)\}}{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta} \{Y_i - m(X_i;\beta)\}} = 0. \tag{7}$$

Substituting the optimal weights into the empirical likelihood in (3), we get

$$L_n(\beta) = \prod_{i=1}^n \frac{1}{n} \frac{1}{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta} \{Y_i - m(X_i;\beta)\}},$$

and the log empirical likelihood is

$$\begin{aligned} \ell_n(\beta) &=: \log\{L_n(\beta)\} \\ &= -\sum_{i=1}^n \log\left\{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta} \{Y_i - m(X_i;\beta)\}\right\} - n\log(n). \end{aligned} \tag{8}$$

The computing intensive nature of the empirical likelihood is clear from the above discussions. Indeed, to evaluate the EL at a $\beta$, one needs to solve the nonlinear equation (7) for the $\lambda$ which depends on $\beta$. An alternative computational approach, as

given in Owen (1990), is to translate the optimization problem (3)–(5) with respect to the EL weights $\{p_i\}_{i=1}^n$ to its dual problem with respect to $\lambda$.

The dual problem to (3)–(5) involves minimizing the objective function

$$Q(\lambda) =: -\sum_{i=1}^{n} \log\left\{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}\{Y_i - m(X_i;\beta)\}\right\},$$

which is the first term in the empirical likelihood ratio in (8), subject to

$$1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}\{Y_i - m(X_i;\beta)\} \geq 1/n \quad \text{for each } i = 1, \ldots, n. \tag{9}$$

The constraint (9) comes from $0 \leq p_i \leq 1$ for each $i$, whereas the gradient of $Q(\lambda)$ is the function on the left-hand side of (7). Let

$$D = \left\{\lambda : 1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}\{Y_i - m(X_i;\beta)\} \geq 1/n \text{ for each } i = 1, \ldots, n\right\}.$$

Then, the dual problem becomes the problem of minimizing $Q(\lambda)$ over the set $D$. It can be verified that $D$ is convex, closed, and compact. Hence, there is a unique minimum within $D$. As suggested in Owen (1990), the set $D$ can be removed by modifying the $\log(x)$ function in $Q(\lambda)$ by a $\log^*(x)$ such that $\log^*(x) = \log(x)$ for $x \geq 1/n$ and $\log^*(x) = -n^2x^2/2 + 2nx - 3/2 - \log(n)$ for $x < 1/n$, which is the quadratic function that matches $\log(x)$ and its first two derivatives at $x = 1/n$.

We note that the profile likelihood $\prod_{i=1}^n p_i$ achieves its maximum $n^{-n}$ when all the weights $p_i$ equal $n^{-1}$ for $i = 1, \ldots, n$. Thus, if there exists a $\beta$, say $\hat{\beta}$, which solves (7) with $\lambda = 0$, namely

$$\sum_{i=1}^{n} \frac{\partial m(X_i;\beta)}{\partial \beta}\{Y_i - m(X_i;\beta)\} = 0, \tag{10}$$

then the EL attains its maximum $L_n(\hat{\beta}) = n^{-n}$ at $\hat{\beta}$. In the parametric regression we are considering, the number of parameters and the number of equations in (10) are the same. Hence, (10) has a solution $\hat{\beta}$ with probability approaching one in large samples. There are inference situations where the number of estimating equations is larger than the number of parameters (strictly speaking, dimension of the parameter space), for instance, the Generalized Method of Moments in econometrics (Hansen 1982). Here, more model information is accounted for by imposing more moment restrictions, leading to more estimating equations than the number of parameters in the model. In statistics, they appear in the form of extra model information. In these so-called over-identified situations, the maximum EL, still using the notation $L_n(\hat{\beta})$, may be different from $n^{-n}$. See Qin and Lawless (1994) for a discussion on this issue.

Following the convention of the standard parametric likelihood, we can define from (8) the log EL ratio

$$r_n(\beta) = -2\log\{L_n(\beta)/L_n(\hat{\beta})\} = 2\sum_{i=1}^{n} \log\left\{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))\right\}. \tag{11}$$

Wilks' theorem (Wilks 1938) is a key property of the parametric likelihood ratio. If we replace the EL $L_n(\beta)$ by the corresponding parametric likelihood, say $L_{pn}(\beta)$, and use $r_{pn}(\beta)$ to denote the parametric likelihood ratio, according to Wilks' theorem, under certain regularity conditions,

$$r_{pn}(\beta_0) \xrightarrow{d} \chi_p^2 \quad \text{as } n \to \infty. \tag{12}$$

This property is maintained by the EL, as is demonstrated in Owen (1990) for the mean parameter, Owen (1991) for linear regression, and many other situations (Qin and Lawless 1994; Molanes-López et al. 2009). In the context of parametric regression,

$$r_n(\beta_0) \xrightarrow{d} \chi_p^2 \quad \text{as } n \to \infty. \tag{13}$$

This can be viewed as a nonparametric version of Wilks' theorem, and it is quite remarkable for the empirical likelihood to achieve such a property under a nonparametric setting with much less parametric distributional assumptions. We call this analogue of sharing the Wilks' theorem the first-order analogue between the parametric and the empirical likelihood.

To appreciate why the nonparametric version of Wilks' theorem is valid, we would like to present a few steps of derivation that offer some insights into the nonparametric likelihood. Typically, the first step in a study on EL is considering an expansion for $\lambda$ defined in (7) at $\beta_0$, the true value of $\beta$, and determining its order of magnitude. It can be shown that for the current parametric regression,

$$\lambda = O_p\big(n^{-1/2}\big). \tag{14}$$

Such a rate for $\lambda$ is obtained in the original papers of Owen (1988, 1990) for the mean parameter (which can be treated as a trivial case of regression without covariates), in Owen (1991) for linear regression, and in Qin and Lawless (1994) and Molanes-López et al. (2009) for the more general case of estimating equations.

With (14), (7) can be inverted (see DiCiccio et al. 1989, for more details). To simplify the notation, define $Z_{ni} = \frac{\partial m(X_i;\beta_0)}{\partial \beta_0}\{Y_i - m(X_i;\beta_0)\}$. Then, (7) can be inverted as

$$n^{-1}\sum_{i=1}^{n} Z_{ni}\big(1 - \lambda^T Z_{ni}\big) + n^{-1}\sum_{i=1}^{n} Z_{ni}\frac{\lambda^T Z_{ni} Z_{ni}^T \lambda}{1 + \lambda^T Z_{ni}} = 0.$$

The last term on the left-hand side (LHS) is $O_p(n^{-1})$, which is negligible relative to the first term on the LHS. Therefore,

$$\lambda = S_n^{-1} n^{-1}\sum_{i=1}^{n} Z_{ni} + o_p\big(n^{-1/2}\big),$$

where $S_n = n^{-1}\sum_{i=1}^{n} Z_{ni} Z_{ni}^T$. Applying a Taylor expansion of $\log(\cdot)$ around 1 and substituting this one-term expansion into the EL ratio $r_n(\beta_0)$ in (11), we have for

some $\gamma_i$ between 1 and $1+\lambda^T Z_{ni}$ $(i=1,\ldots,n)$,

$$\begin{aligned} r_n(\beta_0) &= 2\sum_{i=1}^{n}\log\bigl(1+\lambda^T Z_{ni}\bigr) \\ &= 2\sum_{i=1}^{n}\left\{\lambda^T Z_{ni} - \frac{1}{2}\bigl(\lambda^T Z_{ni}\bigr)^2 + \frac{1}{3}\frac{(\lambda^T Z_{ni})^3}{(1+\gamma_i)^3}\right\} \\ &= 2\lambda^T\sum_{i=1}^{n} Z_{ni} - \lambda^T\sum_{i=1}^{n} Z_{ni}Z_{ni}^T\lambda + O_p\bigl(n^{-1/2}\bigr) \\ &= \left(n^{-1}\sum_{i=1}^{n} Z_{ni}\right)^T S_n^{-1}\left(n^{-1}\sum_{i=1}^{n} Z_{ni}\right) + o_p(1), \end{aligned} \tag{15}$$

which leads to Wilks' theorem as $S_n \xrightarrow{p} \Sigma(\beta_0) =: E\{Z_{ni}Z_{ni}^T\}$ and

$$n^{-1/2}\sum_{i=1}^{n} Z_{ni} \xrightarrow{d} N\bigl(0,\Sigma(\beta_0)\bigr) \quad \text{as } n\to\infty.$$

As commonly practiced in parametric likelihood, the above nonparametric version of Wilks' theorem can be used to construct likelihood ratio confidence regions for $\beta_0$. An EL confidence region with a nominal level of confidence $1-\alpha$ is

$$I_{1-\alpha} = \bigl\{\beta : r_n(\beta) \le \chi^2_{p,1-\alpha}\bigr\},$$

where $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$-quantile of the $\chi^2_p$ distribution. Wilks' theorem in (13) ensures that

$$P\{\beta_0 \in I_{1-\alpha}\} \to 1-\alpha \quad \text{as } n\to\infty.$$

This construction mirrors the conventional likelihood ratio confidence regions except that the EL ratio is employed here instead of the parametric likelihood ratio.

Note that (15) also shows that the EL method is (first-order) asymptotically equivalent to the normal approximation method. However, the normal method requires the estimation of the variance $\Sigma(\beta_0)$, whereas the EL method does not require any explicit variance estimation. This is because the studentization is carried out internally via the optimization procedure.

In addition to the first-order analogue between the parametric and the empirical likelihood, there is a second-order analogue between them in the form of the Bartlett correction. Bartlett correction is an elegant second-order property of the parametric likelihood ratios, which was conjectured and proposed in Bartlett (1937). It was formally established and studied in a series of papers including Lawley (1956), Hayakawa (1977), Barndorff-Nielsen and Cox (1984), and Barndorff-Nielsen and Hall (1988).

Let $w_i = \Sigma(\beta_0)^{-1/2}Z_{ni} = (w_i^1,\ldots,w_i^p)^T$ and for $j_l \in \{1,\ldots,p\}$, $l=1,\ldots,k$, define $\alpha^{j_1\cdots j_k} = E(w_i^{j_1}\cdots w_i^{j_k})$ for a $k$th multivariate cross moments of $w_i$. By as-

suming the existence of higher-order moments of $Z_{ni}$, it may be shown via developing Edgeworth expansions that the distribution of the empirical likelihood ratio admits the following expansion:

$$P\{r_n(\beta_0) \le \chi^2_{p,1-\alpha}\} = 1 - \alpha - a\,\chi^2_{p,1-\alpha} g_p\big(\chi^2_{p,1-\alpha}\big)\, n^{-1} + O\big(n^{-3/2}\big), \tag{16}$$

where $g_p$ is the density of the $\chi^2_p$ distribution, and

$$a = p^{-1}\left(\frac{1}{2}\sum_{j,m=1}^{p} \alpha^{jjmm} - \frac{1}{3}\sum_{j,k,m=1}^{p} \alpha^{jkm}\alpha^{jkm}\right). \tag{17}$$

This means that for the parametric regression, both parametric and empirical likelihood ratio confidence regions $I_{1-\alpha}$ have coverage error of order $n^{-1}$. Part of the coverage error is due to the fact that the mean of $r_n(\beta_0)$ does not agree with $p$, the mean of $\chi^2_p$, that is, $E\{r_n(\beta_0)\} \neq p$, but rather

$$E\{r_n(\beta_0)\} = p\big(1 + an^{-1}\big) + O\big(n^{-2}\big),$$

where $a$ has been given above.

The idea of the Bartlett correction is to adjust the EL ratio $r_n(\beta_0)$ to $r^*_n(\beta_0) = r_n(\beta_0)/(1 + an^{-1})$ so that $E\{r^*_n(\beta_0)\} = p + O(n^{-2})$. And amazingly this simple adjustment to the mean leads to improvement in (16) by one order of magnitude (DiCiccio et al. 1991; Chen 1993; and Chen and Cui 2007) so that

$$P\{r^*_n(\beta_0) \le \chi^2_{p,1-\alpha}\} = 1 - \alpha + O\big(n^{-2}\big). \tag{18}$$

## 3 Nonparametric regression

Consider in this section the nonparametric regression model

$$Y_i = m(X_i) + \varepsilon_i, \tag{19}$$

where the regression function $m(x) = E(Y_i|X_i = x)$ is nonparametric, and $X_i$ is $d$-dimensional. We assume that the regression can be heteroscedastic in that $\sigma^2(x) = \mathrm{Var}(Y_i|X_i = x)$, the conditional variance of $Y_i$ given $X_i = x$, may depend on $x$.

The kernel smoothing method is a popular method for estimating $m(x)$ nonparametrically. See Härdle (1990) and Fan and Gijbels (1996) for comprehensive overviews. Other nonparametric methods for estimating $m(x)$ include splines, orthogonal series, and wavelets methods. The simplest kernel regression estimator for $m(x)$ is the following Nadaraya–Watson estimator:

$$\hat{m}(x) = \frac{\sum_{i=1}^n K_h(x - X_i)Y_i}{\sum_{i=1}^n K_h(x - X_i)}, \tag{20}$$

where $K_h(t) = K(t/h)/h^d$, $K$ is a $d$-dimensional kernel function, and $h$ is a bandwidth. The above kernel estimator can be obtained by minimizing the following locally weighted sum of least squares:

$$\sum_{i=1}^{n} K_h(x - X_i)\{Y_i - m(x)\}^2$$

with respect to $m(x)$. It is effectively the solution of the following estimating equation:

$$\sum_{i=1}^{n} K_h(x - X_i)\{Y_i - m(x)\} = 0. \tag{21}$$

Under the nonparametric regression model, the unknown "parameter" is the regression function $m(x)$ itself. The empirical likelihood for $m(x)$ at a fixed $x$ can be formulated in a fashion similar to the parametric regression setting considered in the previous section. Alternatively, since the empirical likelihood is being applied to the weighted average $\sum_{i=1}^{n} K_h(x - X_i)m(x)$, it is also similar to the EL of a mean.

Let $p_1, \ldots, p_n$ be probability weights adding to one. The empirical likelihood evaluated at $\theta(x)$, a candidate value of $m(x)$, is

$$L_n\{\theta(x)\} = \max \prod_{i=1}^{n} p_i, \tag{22}$$

where the maximization is subject to $\sum_{i=1}^{n} p_i = 1$ and

$$\sum_{i=1}^{n} p_i K_h(x - X_i)\{Y_i - \theta(x)\} = 0. \tag{23}$$

By comparing this formulation of the EL with that for the parametric regression, we see that the two formulations are largely similar except that (23) is used as the structural constraint instead of (5). This comparison does highlight the role played by the structural constraint in the EL formulation. Indeed, different structural constraints give rise to EL for different "parameters" (quantity of interest), just like different densities give rise to different parametric likelihoods. In general, the empirical likelihood is formulated based on the parameters of interest via the structural constraints, and the parametric likelihood is fully based on a parametric model.

The algorithm for solving the above optimization problem (22)–(23) is similar to the EL algorithm for the parametric regression given under (4) and (5), except that it may be viewed easier as the "parameter" is one-dimensional if we ignore the issue of bandwidth selection for nonparametric regression. By introducing Lagrange multipliers like we did in (6) in the previous section, we have that the optimal EL weights for the above optimization problem at $\theta(x)$ are given by

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda(x)K_h(x - X_i)\{Y_i - \theta(x)\}},$$

where $\lambda(x)$ is a univariate Lagrange multiplier that satisfies

$$\sum_{i=1}^{n} \frac{K_h(x - X_i)\{Y_i - \theta(x)\}}{1 + \lambda(x)K_h(x - X_i)\{Y_i - \theta(x)\}} = 0. \tag{24}$$

Substituting the optimal weights into the empirical likelihood in (22), the empirical likelihood evaluated at $\theta(x)$ is

$$L_n\big\{\theta(x)\big\} = \prod_{i=1}^{n} \frac{1}{n} \frac{1}{1 + \lambda(x)K_h(x - X_i)\{Y_i - \theta(x)\}},$$

and the log empirical likelihood is

$$\begin{aligned} \ell_n\big\{\theta(x)\big\} &=: \log\big\{L_n\big\{\theta(x)\big\}\big\} \\ &= -\sum_{i=1}^{n} \log\big[1 + \lambda(x)K_h(x - X_i)\big\{Y_i - \theta(x)\big\}\big] - n\log(n). \end{aligned} \tag{25}$$

The overall EL is maximized at $p_i = n^{-1}$, which corresponds to $\theta(x)$ being the Nadaraya–Watson estimator $\hat{m}(x)$ in (20). Hence, we can define the log EL ratio at $\theta(x)$ as

$$\begin{aligned} r_n\big\{\theta(x)\big\} &= -2\log\big[L_n\big\{\theta(x)\big\}/n^{-n}\big] \\ &= 2\sum_{i=1}^{n} \log\big[1 + \lambda(x)K_h(x - X_i)\big\{Y_i - \theta(x)\big\}\big]. \end{aligned} \tag{26}$$

The above EL is not actually for $m(x)$, the true underlying function value at $x$, but rather for $E\{\hat{m}(x)\}$. This can be actually detected by the form of the structural constraint (23). It is well known in kernel estimation that $\hat{m}(x)$ is not an unbiased estimator of $m(x)$, as is the case for almost all nonparametric estimators. For the Nadaraya–Watson estimator,

$$E\big\{\hat{m}(x)\big\} = m(x) + b(x) + o\big(h^2\big),$$

where $b(x) = \frac{1}{2}h^2\{m''(x) + 2m'(x)f'(x)/f(x)\}$ is the leading bias of the kernel estimator, and $f$ is the density of $X_i$. Then, the EL is actually evaluated at a $\theta(x)$, that is a candidate value of $m(x) + b(x)$ instead of $m(x)$. There are two strategies to reduce the effect of the bias (Hall 1991). One is to undersmooth with a bandwidth $h = o(n^{-1/(4+d)})$, the optimal order of bandwidth that minimizes the mean squared error of estimation with a second-order kernel ($d$ is the dimension of $X$). Another is to explicitly estimate the bias and then to subtract it from the kernel estimate. We consider the first approach of undersmoothing here for reasons of simplicity.

When undersmoothing so that $n^{2/(4+d)}h^2 \to 0$, Wilks' theorem is valid for the EL under the current nonparametric regression in that

$$r_n\big\{m(x)\big\} \xrightarrow{d} \chi_1^2 \quad \text{as } n \to \infty.$$

This means that an empirical likelihood confidence interval with nominal coverage equal to $1-\alpha$, denoted as $I_{1-\alpha,el}$, is given by

$$I_{1-\alpha,el} = \{\theta(x) : r_n\{\theta(x)\} \le \chi^2_{1,1-\alpha}\}. \tag{27}$$

A special feature of the empirical likelihood confidence interval is that no explicit variance estimator is required in its construction as the studentization is carried out internally via the optimization procedure.

Define $\omega_i = K_h(x - X_i)\{Y_i - m(x)\}$ and, for positive integers $j$,

$$\bar{\omega}_j = n^{-1}\sum_{i=1}^{n}\omega_i^j, \qquad \mu_j = E(\bar{\omega}_j) \quad \text{and} \quad R_j(K) = \int K^j(u)\,du.$$

We note here that the bias in the kernel smoothing makes $\mu_1 = O(h^2)$, while in the parametric regression case $\mu_1 = 0$.

It is shown in Chen and Qin (2003) that the coverage probability of $I_{1-\alpha,el}$ admits the following Edgeworth expansion:

$$\begin{aligned} &P\{m(x) \in I_{1-\alpha,el}\} \\ &\quad = 1-\alpha - \left\{nh^d\mu_1^2\mu_2^{-1} + \left(\frac{1}{2}\mu_2^{-2}\mu_4 - \frac{1}{3}\mu_2^{-3}\mu_3^2\right)(nh^d)^{-1}\right\} z_{1-\frac{\alpha}{2}}\phi(z_{1-\frac{\alpha}{2}}) \\ &\qquad + O\{nh^{d+6} + h^4 + (nh^d)^{-1}h^2 + (nh^d)^{-2}\}, \end{aligned} \tag{28}$$

where $\phi$ and $z_{1-\frac{\alpha}{2}}$ are the density and the $(1-\frac{\alpha}{2})$-quantile of a standard normal random variable.

The above expansion is nonstandard in that the leading coverage error consists of two terms. The first term, $nh^d\mu_1\mu_2^{-1}$, of order $nh^{d+4}$ is due to the bias in the kernel smoothing. The second term of order $(nh^d)^{-1}$ is largely similar to the leading coverage error for parametric regression in (16). We note that in the second term, the effective sample size in the nonparametric estimation near $x$ is $nh^d$ instead of $n$, the effective sample size in the parametric regression.

The next question is if the Bartlett correction is still valid under the nonparametric regression. The answer is yes. It may be shown that

$$E[r_n\{m(x)\}] = 1 + (nh^d)^{-1}\gamma + o\{nh^{d+4} + (nh^d)^{-1}\},$$

where

$$\gamma = \mu_2^{-1}(nh^d\mu_1)^2 + \frac{1}{2}\mu_2^{-2}\mu_4 - \frac{1}{3}\mu_2^{-3}\mu_3^2. \tag{29}$$

Note that $\gamma$ appears in the leading coverage error term in (28). Based on (28) and choosing $h = O(n^{-\frac{1}{d+2}})$, we have, with $c_\alpha = \chi^2_{1,1-\alpha}$,

$$\begin{aligned} &P[r_n\{m(x)\} \le c_\alpha\{1 + \gamma(nh^d)^{-1}\}] \\ &\quad = P[\chi_1^2 \le c_\alpha\{1 + \gamma(nh^d)^{-1}\}] \end{aligned}$$

$$
\begin{aligned}
&- \left(nh^d\right)^{-1}\gamma c_\alpha^{1/2}\left\{1+\gamma\left(nh^d\right)^{-1}\right\}^{1/2}\phi\left[c_\alpha^{-1/2}\left\{1+\gamma\left(nh^d\right)^{-1}\right\}^{1/2}\right] \\
&+ O\left\{\left(nh^d\right)^{-2}\right\} \\
&= P\left(\chi_1^2 \le c_\alpha\right) + \left(nh^d\right)^{-1}\gamma z_{1-\frac{\alpha}{2}}\phi(z_{1-\frac{\alpha}{2}}) - \left(nh^d\right)^{-1}\gamma z_{1-\frac{\alpha}{2}}\phi(z_{1-\frac{\alpha}{2}}) \\
&+ O\left\{\left(nh^d\right)^{-2}\right\} \\
&= 1-\alpha + O\left(n^{-\frac{4}{d+2}}\right). \qquad (30)
\end{aligned}
$$

Therefore, the empirical likelihood is Bartlett correctable in the current context of nonparametric regression. In practice, the Bartlett factor $\gamma$ has to be estimated, say by a consistent $\hat{\gamma}$. Chen and Qin (2003) gave more details on practical implementation; see also Chen (1996) for an implementation in the case of density estimation.

## 4 Semiparametric regression

We next consider the empirical likelihood method in the context of semiparametric regression.

### 4.1 Partial linear regression model

Let us first consider the partial linear model, defined as follows:

$$Y_i = \beta^T X_i + g(Z_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n, \qquad (31)$$

where the response $Y_i$ and the explanatory variable $Z_i$ are one-dimensional, $\beta$ and $X_i$ are $p$-dimensional ($p \ge 1$), and $g(\cdot)$ is a continuous but unknown nuisance function. It is assumed that the error $\varepsilon_i$ satisfies $E(\varepsilon_i | X_i, Z_i) = 0$ and $\mathrm{Var}(\varepsilon_i | X_i, Z_i) = \sigma^2$.

Our goal is to construct confidence regions or test hypotheses concerning the vector $\beta_0$ of true regression coefficients. For this, we first need to estimate the unknown function $g$. Define for fixed $\beta$,

$$\hat{g}_\beta(z) = \sum_{i=1}^n \frac{K_h(z - Z_i)}{\sum_{j=1}^n K_h(z - Z_j)}\left(Y_i - \beta^T X_i\right),$$

where $K_h(u) = K(u/h)/h$, $h = h_n$ is a smoothing parameter, and $K$ is a (one-dimensional) kernel function (probability density function). Instead of the above local constant estimator, we could also use, e.g., local polynomial estimators. The idea is now to mimic the empirical likelihood method developed for parametric regression, but considering $Y - \hat{g}_\beta(Z)$ as the new (artificial) response. This leads to the following likelihood ratio function:

$$R_n(\beta) = \max \prod_{i=1}^n (np_i),$$

where the maximum is taken over all $n$-tuples $(p_1, \ldots, p_n)$ that satisfy

$$p_i \geq 0 \quad (i = 1, \ldots, n), \qquad \sum_{i=1}^{n} p_i = 1,$$

$$\sum_{i=1}^{n} p_i \left\{ X_i + \frac{\partial \hat{g}_\beta(Z_i)}{\partial \beta} \right\} \big( Y_i - \beta^T X_i - \hat{g}_\beta(Z_i) \big) = 0.$$

Note that the latter constraint is equivalent to

$$\sum_{i=1}^{n} p_i \tilde{X}_i \big( \tilde{Y}_i - \beta^T \tilde{X}_i \big) = 0,$$

where

$$\tilde{X}_i = X_i - \sum_{j=1}^{n} \frac{K_h(Z_i - Z_j)}{\sum_{k=1}^{n} K_h(Z_i - Z_k)} X_j \quad \text{and} \quad \tilde{Y}_i = Y_i - \sum_{j=1}^{n} \frac{K_h(Z_i - Z_j)}{\sum_{k=1}^{n} K_h(Z_i - Z_k)} Y_j$$

are estimators of $X_i - E(X|Z = Z_i)$ and $Y_i - E(Y|Z = Z_i)$, respectively. Wang and Jing (2003) showed that under certain regularity conditions, the following result holds:

$$r_n(\beta_0) = -2 \log R_n(\beta_0) \xrightarrow{d} \chi^2_p.$$

This result shows that asymptotically, the estimation of the unknown function $g$ has no influence on the asymptotic limit, as we get exactly the same limit as in the parametric case, i.e., as in the case where the function $g$ would be known. This result is important, as it shows that we can obtain empirical likelihood confidence regions for $\beta_0$ without estimating any variance.

When the interest lies in testing the validity of the whole partial linear model by means of an EL approach (instead of testing only the value of the parameter vector $\beta_0$), one can use the method developed by Chen and Van Keilegom (2009) and Van Keilegom et al. (2008). In the former paper the authors developed a general smoothing based EL approach to test the validity of any semiparametric model, whereas the latter paper considers the same testing problem, but by using an EL approach based on marked empirical processes, which is quite different in nature from the former approach. See also Sect. 7 for more details.

### 4.2 Single-index regression model

Let us now consider the case of single-index models. Suppose that the relation between the (one-dimensional) response $Y_i$ and the $p$-dimensional vector $X_i$ of explanatory variables is given by

$$Y_i = g\big(\beta^T X_i\big) + \varepsilon_i, \tag{32}$$

where $g$ is an unknown but smooth nuisance function, and the error $\varepsilon_i$ satisfies $E(\varepsilon_i|X_i) = 0$ and $\mathrm{Var}(\varepsilon_i|X_i) = \sigma^2$. Let $\beta_0$ be the true parameter vector. In order

to identify the model, we suppose that $\|\beta\| = 1$, where $\|\cdot\|$ denotes the Euclidean norm. For any $\beta = (\beta_1, \ldots, \beta_p)^T$ satisfying $\|\beta\| = 1$ and any $1 \le r \le p$, let $\beta^{(r)} = (\beta_1, \ldots, \beta_{r-1}, \beta_{r+1}, \ldots, \beta_p)^T$, and supposing that $\beta_r > 0$, we can write $\beta = (\beta_1, \ldots, \beta_{r-1}, (1 - \|\beta^{(r)}\|^2)^{1/2}, \beta_{r+1}, \ldots, \beta_p)^T$. Now, let $J_{\beta^{(r)}}$ be the $p \times (p-1)$ Jacobian matrix given by

$$J_{\beta^{(r)}} = \frac{\partial \beta}{\partial \beta^{(r)}} = (\gamma_1, \ldots, \gamma_p)^T$$

with $\gamma_s$ $(s \neq r)$ the unit vector with $s$th component equal to one, and $\gamma_r = -(1 - \|\beta^{(r)}\|^2)^{-1/2}\beta^{(r)}$. Now, it can be easily seen that $E[\xi_i(\beta_0^{(r)})] = 0$ $(i = 1, \ldots, n)$, where

$$\xi_i\big(\beta^{(r)}\big) = \big[Y_i - g\big(\beta^T X_i\big)\big] g'\big(\beta^T X_i\big) J_{\beta^{(r)}}^T X_i.$$

Hence, it seems natural to use the $\xi_i(\beta^{(r)})$'s as building blocks of the empirical likelihood ratio. However, since they depend on the unknown functions $g$ and $g'$, we first replace them by suitable estimators. Let

$$\hat{g}(t;\beta) = \sum_{i=1}^n \frac{W_{ni}(t;\beta,h)Y_i}{\sum_{j=1}^n W_{nj}(t;\beta,h)},$$

$$\hat{g}'(t;\beta) = \sum_{i=1}^n \frac{\tilde{W}_{ni}(t;\beta,h)Y_i}{\sum_{j=1}^n W_{nj}(t;\beta,h)}$$

be local linear estimators of $g(t)$ and $g'(t)$, where $W_{ni}(t;\beta,h) = K_h(\beta^T X_i - t) \times [S_{n2}(t;\beta,h) - (\beta^T X_i - t)S_{n1}(t;\beta,h)]$, $\tilde{W}_{ni}(t;\beta,h) = K_h(\beta^T X_i - t)[(\beta^T X_i - t) \times S_{n0}(t;\beta,h) - S_{n1}(t;\beta,h)]$, and $S_{nk}(t;\beta,h) = n^{-1}\sum_{i=1}^n (\beta^T X_i - t)^k K_h(\beta^T X_i - t)$ $(k = 0, 1, 2)$. We are now ready to define the empirical likelihood ratio. Define

$$R_n(\beta^{(r)}) = \max \prod_{i=1}^n (np_i),$$

where the maximum is taken over all $(p_1, \ldots, p_n)$ that satisfy

$$p_i \ge 0 \quad (i = 1, \ldots, n), \qquad \sum_{i=1}^n p_i = 1, \qquad \sum_{i=1}^n p_i \hat{\xi}_i\big(\beta^{(r)}\big) = 0,$$

where

$$\hat{\xi}_i\big(\beta^{(r)}\big) = \big[Y_i - \hat{g}\big(\beta^T X_i;\beta\big)\big]\hat{g}'\big(\beta^T X_i;\beta\big) J_{\beta^{(r)}}^T X_i.$$

Then, Xue and Zhu (2006) showed that under suitable regularity conditions,

$$-2\log R_n\big(\beta_0^{(r)}\big) \xrightarrow{d} w_1\chi_{1,1}^2 + \cdots + w_{p-1}\chi_{1,p-1}^2$$

for certain weights $w_1, \ldots, w_{p-1}$, where $\chi_{1,1}^2, \ldots, \chi_{1,p-1}^2$ are independent $\chi_1^2$ variables. Hence, Wilks' theorem is not valid here. Since the weights are unknown, they

need to be replaced by suitable estimators, before we can apply the above limit to construct confidence regions for the vector $\beta_0^{(r)}$.

In order to circumvent this problem, one can also redefine the empirical likelihood in the following way. Instead of working with the $\hat{\xi}_i(\beta^{(r)})$'s, we build the empirical likelihood with the following adjusted quantities:

$$\hat{\eta}_i\big(\beta^{(r)}\big) = \big[Y_i - \hat{g}\big(\beta^T X_i; \beta\big)\big]\hat{g}'\big(\beta^T X_i; \beta\big) J_{\beta^{(r)}}^T \big[X_i - \hat{E}\big(X_i|\beta^T X_i\big)\big],$$

where

$$\hat{E}\big(X_i|\beta^T X_i = t\big) = \sum_{i=1}^{n} \frac{W_{ni}(t; \beta, h) X_i}{\sum_{j=1}^{n} W_{nj}(t; \beta, h)}.$$

Now, let $\tilde{R}_n(\beta^{(r)})$ be the EL ratio obtained by replacing the $\hat{\xi}_i$'s by $\hat{\eta}_i$'s. Then, Zhu and Xue (2006) showed that Wilks' theorem holds, i.e.,

$$-2\log \tilde{R}_n\big(\beta_0^{(r)}\big) \xrightarrow{d} \chi^2_{p-1}.$$

They also showed a similar result in the case where the model is a so-called partially linear single-index model, i.e., where the regression function is the sum of a linear component and a single-index component.

As for the partial linear model, the validity of the single-index model can be tested by using the tests developed by Chen and Van Keilegom (2009) and Van Keilegom et al. (2008). The above asymptotic results can also be obtained from Hjort et al. (2009), who developed generic conditions for the asymptotic theory of any EL ratio, built up using estimating equations depending on plug-in estimators of unknown nuisance parameters (see also Sect. 6.3).

## 5 Regression with missing values

Often in statistical applications, the data collected for a regression analysis, say $\{(X_1^T, Y_1), \ldots, (X_n^T, Y_n)\}^T$, contain missing values. The missing values can be either in the responses $Y_i$ or in the covariates $X_i$. However, we do not allow any component of $Y_i$ or $X_i$ to be always missing, namely we rule out the case where some components of the data are completely latent.

We start with the easier case of missing responses, and then we discuss the missing covariates.

### 5.1 Missing responses

Assume the parametric regression model (1), given by $Y_i = m(X_i; \beta) + \varepsilon_i$, where $Y_i$ is one-dimensional, $X_i$ is $d$-dimensional, and $\beta$ is $p$-dimensional, and assume that the data $(X_i^T, Y_i)^T$ $(i = 1, \ldots, n)$ are i.i.d. Due to nonresponse or other reasons in the data collection, $Y_i$ is subject to missingness. Here we assume that all $X_i$'s are always observed.

Let $\delta_i$ be the missing indicator of $Y_i$ such that $\delta_i = 0$ (1) for missing (observed) $Y_i$. The data we observe can be expressed as

$$\{(X_i, Y_i \delta_i)\}_{i=1}^n.$$

The Strongly Ignorable Missing at Random mechanism (MAR) (Rubin 1976, and Rosenbaum and Rubin 1983) is an important notion in missing data analysis. In the case of missing responses, MAR means that the missingness of $Y_i$ is predictable by the observable covariate $X_i$ so that conditioning on the covariate $X_i$, the missingness of $Y_i$ is independent of $Y_i$ itself. Put in mathematical terms,

$$P(\delta_i = 1|Y_i, X_i) = P(\delta_i = 1|X_i) =: w(X_i). \tag{33}$$

Here, $w$ is called the missing propensity of $Y_i$. A stronger form of missingness than MAR is the so-called Missing Completely at Random (MCAR) since the latter implies that the propensity $w(x)$ is a constant function.

When the missingness is MAR but not MCAR, there is a selection bias in the mechanism that generates the missing values. In this case, simply deleting missing values will produce biased estimators and misleading inference.

Suppose that we have a parametric model for the missing propensity function $w(x; \theta)$ where $\theta$ is a $q$-dimensional parameter. Without too much abuse of notation, let $f$ denote a generic probability "density" function. Here "density" should be interpreted in a general sense with respect to the probability measure. Under the MAR, the full likelihood of the observed data is

$$\begin{aligned}
\mathcal{L}_n &= \prod_{\delta_i=1} f(X_i, Y_i, \delta_i = 1) \prod_{\delta_i=0} f(X_i, \delta_i = 0) \\
&= \prod_{\delta_i=1} f(X_i, Y_i) p(\delta_i = 1|X_i, Y_i) \prod_{\delta_i=0} f(X_i) p(\delta_i = 0|X_i) \\
&= \prod_{i=1}^n w(X_i; \theta)^{\delta_i} \{1 - w(X_i; \theta)\}^{1-\delta_i} \prod_{\delta_i=1} f(X_i, Y_i) \prod_{\delta_i=0} f(X_i).
\end{aligned} \tag{34}$$

We have not specified the parameters that define the "densities" of $(X, Y)$ and $X$ since doing so is not important for our inference for regression. Let

$$L_B(\theta) = \prod_{i=1}^n w(X_i; \theta)^{\delta_i} \{1 - w(X_i; \theta)\}^{1-\delta_i}$$

be the binary likelihood associated with the missing mechanism. It is reasonable to assume that the missing propensity parameter $\theta$ is not involved in defining the just mentioned "densities" $f$. In this case, the likelihood $\mathcal{L}_n$ in (34) can be partitioned into two parts, one purely for $\theta$ and the other for the parameters that define the joint density of $(X_i, Y_i)$. Hence, $\theta$ can be estimated by maximizing the binary likelihood $L_B(\theta)$. Let us denote this estimator by $\hat{\theta}$, i.e.,

$$\hat{\theta} = \arg\max_\theta L_B(\theta). \tag{35}$$

A simple estimator of $\beta$ is the so-called complete-case-based estimator. Define the least square function of $\beta$:

$$LS_c(\beta) = \sum_{i=1}^{n} \delta_i \{Y_i - m(X_i; \beta)\}^2.$$

Minimizing $LS_c(\beta)$ leads to a complete-case-based LS estimator $\hat{\beta}_c$ which is the solution of

$$\sum_{i=1}^{n} \delta_i \frac{\partial m(X_i; \beta)}{\partial \beta} \{Y_i - m(X_i; \beta)\} = 0. \tag{36}$$

The empirical likelihood for $\beta$ can be constructed analogously to the formulation from (4) to (5) without missing values. Specifically, the EL for $\beta$ is

$$L_{nc}(\beta) = \max \prod_{i=1}^{n} p_i \tag{37}$$

subject to

$$\sum_{i=1}^{n} p_i \delta_i = 1 \quad \text{and} \tag{38}$$

$$\sum_{i=1}^{n} p_i \delta_i \frac{\partial m(X_i; \beta)}{\partial \beta} \{Y_i - m(X_i; \beta)\} = 0. \tag{39}$$

Let $r_{nc}(\beta) = -2\log\{L_{nc}(\beta)/n^{-n}\}$ be the log EL ratio. It can be shown that both Wilks' theorem and the Bartlett correction are valid in this case of missing values.

Another approach for constructing EL in the case of missing values is based on the notion of imputation. Given the consistent estimator $\hat{\beta}_c$, we impute a missing $Y_i$ by $Y_i^* = m(X_i; \hat{\beta}_c)$. The EL for $\beta$ can be formed by

$$L_{nI}(\beta) = \max \prod_{i=1}^{n} p_i \tag{40}$$

subject to $\sum_{i=1}^{n} p_i = 1$ and

$$\sum_{i=1}^{n} p_i \frac{\partial m(X_i; \beta)}{\partial \beta} \big[\{Y_i - m(X_i; \beta)\}\delta_i + \{Y_i^* - m(X_i; \beta)\}(1 - \delta_i)\big] = 0. \tag{41}$$

The above EL formulation can be extended to other parameters. For instance, if our interest is on inference for the marginal mean of $Y$, say $\mu_y = E(Y)$, Wang and Rao (2002) proposed the following EL for $\mu_y$:

$$L_n(\mu_y) = \max \prod_{i=1}^{n} p_i \tag{42}$$

subject to $\sum p_i = 1$ and

$$\sum_{i=1}^{n} p_i \{Y_i \delta_i + Y_i^*(1-\delta_i) - \mu_y\} = 0.$$

Due to using the imputed values, the EL ratio statistic may not admit Wilks' theorem.

When the regression function is nonparametric as specified in (19) instead of parametric, both the complete-case-based method and the imputation method outlined above for parametric regression can be extended to nonparametric regression.

The complete-case-based empirical likelihood evaluated at $\theta(x)$, a candidate value of $m(x)$, is

$$L_{nc}\{\theta(x)\} = \max \prod_{i=1}^{n} p_i$$

subject to $\sum_{i=1}^{n} p_i = 1$ and

$$\sum_{i=1}^{n} p_i \delta_i K_h(x - X_i)\{Y_i - \theta(x)\} = 0. \tag{43}$$

The nonparametric imputation of a missing $Y_i$ can be achieved by $Y_i^* = \hat{m}_c(x)$, where

$$\hat{m}_c(x) = \frac{\sum_{i=1}^{n} \delta_i K_h(x - X_i) Y_i}{\sum_{i=1}^{n} \delta_i K_h(x - X_i)}.$$

An imputation-based EL for $m(x)$ is

$$L_{nI}\{m(x)\} = \max \prod_{i=1}^{n} p_i$$

subject to $\sum_{i=1}^{n} p_i = 1$ and

$$\sum_{i=1}^{n} p_i K_h(x - X_i)\big[\{Y_i - \theta(x)\}\delta_i + \{Y_i^* - \theta(x)\}(1-\delta_i)\big] = 0.$$

It can be shown that the complete-case-based EL for $m(x)$ will still enjoy Wilks' theorem and the Bartlett correction. However, the imputation-based EL may not be so due to the fact that the imputed $Y_i^*$ does not have the same distribution as the original $Y_i$. Despite this, the imputed EL confidence regions will be smaller than those based on the complete-case EL ratio, which is not surprising since the latter regions are not using all the information available in the data.

### 5.2 Missing covariates

A more challenging type of missing values is missing covariates where the covariate $X_i$ is subject to missingness.

Let $X_i^T = (X_i^{(1)T}, X_i^{(2)T})$ be a partition of $X_i$, where $X_i^{(l)}$ is of dimension $d_l$ ($l = 1, 2$), and $d = d_1 + d_2$. Without loss of generality, we assume that $X_i^{(1)}$ is subject to missingness, whereas $X_i^{(2)}$ and $Y_i$ are always observable.

Redefine $\delta_i = 1$ (0) if $X_i^{(1)}$ is observed (missing). The MAR mechanism becomes

$$P(\delta_i = 1|X_i, Y_i) = P\big(\delta_i = 1|X_i^{(2)}, Y_i\big) =: w_2\big(X_i^{(2)}, Y_i\big).$$

For parametric regression, the complete-case estimation that ignores missing values is attained by minimizing

$$\sum_{i=1}^{n} \delta_i \big\{Y_i - m(X_i; \beta)\big\}^2$$

with respect to $\beta$, which is the same as (36). And, both the estimator for $\beta$ and the EL formulation are the same as those given in (37)–(39). This means that Wilks' theorem and the Bartlett correction will be maintained for the EL in this case.

However, unlike the missing response case, the parametric imputation approach is not straightforward to be implemented as the parametric regression does not specify the conditional distribution of $X_i^{(1)}$ given $(X_i^{(2)}, Y_i)$. If we assume a parametric model for the missing propensity, say $w_2(X_i^{(2)}, Y_i; \theta)$, a more efficient formulation can be achieved by inversely weighting the complete cases. Here, the efficiency means the size of the confidence regions. In this case, the weighted least square function is

$$\sum_{i=1}^{n} \delta_i w_2^{-1}\big(X_i^{(2)}, Y_i; \hat{\theta}\big)\big\{Y_i - m(X_i; \beta)\big\}^2,$$

where $\hat{\theta}$ is the binary likelihood estimator which can be constructed in a similar fashion to (35), and provided that $w_2$ is uniformly bounded away from 0.

The EL for $\beta$ is

$$L_{n2}(\beta) = \max \prod_{i=1}^{n} p_i$$

subject to $\sum_{i=1}^{n} p_i = 1$ and

$$\sum_{i=1}^{n} p_i \delta_i w_2^{-1}\big(X_i^{(2)}, Y_i; \hat{\theta}\big) \frac{\partial m(X_i; \beta)}{\partial \beta}\big\{Y_i - m(X_i; \beta)\big\} = 0.$$

The use of $\hat{\theta}$ can alter the standard asymptotic properties of the EL. To appreciate this point, let

$$Z_i(\beta, \hat{\theta}) = \delta_i w_2^{-1}\big(X_i^{(2)}, Y_i; \hat{\theta}\big) \frac{\partial m(X_i; \beta)}{\partial \beta}\big\{Y_i - m(X_i; \beta)\big\}.$$

Then, according to the EL algorithm as outlined earlier, the log EL ratio equals

$$r_{n2}(\beta) = -2\log\{L_n(\beta)/n^{-n}\} = 2\sum_{i=1}^{n}\log\{1+\lambda^T Z_i(\beta,\hat{\theta})\},$$

where $\lambda$ satisfies

$$n^{-1}\sum_{i=1}^{n}\frac{Z_i(\beta,\hat{\theta})}{1+\lambda^T Z_i(\beta,\hat{\theta})} = 0. \tag{44}$$

By carrying out expansions for $\lambda$ first and then substituting these expansions into $r_{n2}(\beta)$, we have

$$r_{n2}(\beta_0) = n\bar{Z}_n^T(\beta_0,\hat{\theta})S_n^{-1}(\beta_0,\hat{\theta})\bar{Z}_n(\beta_0,\hat{\theta}) + o_p(1),$$

where $\bar{Z}_n(\beta_0,\hat{\theta}) = n^{-1}\sum_{i=1}^{n} Z_i(\beta_0,\hat{\theta})$ and $S_n(\beta_0,\hat{\theta}) = n^{-1}\sum_{i=1}^{n} Z_i(\beta_0,\hat{\theta}) \times Z_i^T(\beta_0,\hat{\theta})$. As $\hat{\theta}$ is $\sqrt{n}$-consistent to $\theta_0$, $S_n(\beta_0,\hat{\theta}) \xrightarrow{p} \Sigma(\beta_0,\theta_0) =: E\{Z_i(\beta_0,\theta_0) \times Z_i^T(\beta_0,\theta_0)\}$. If $\bar{Z}_n(\beta_0,\hat{\theta})$ were asymptotically normal with mean zero and variance $\Sigma(\beta_0,\theta_0)$, then the log EL ratio would be asymptotically chi-squared, and hence Wilks' theorem would be valid. However, due to the use of the estimator $\hat{\theta}$, $\bar{Z}_n(\beta_0,\hat{\theta})$ is asymptotically normal with mean zero but a variance that is different from $\Sigma(\beta_0,\theta_0)$. Hence, the log EL ratio no longer satisfies Wilks' theorem; rather it will be distributed as $\sum_{l=1}^{p} c_l\chi_{1,l}^2$, where $\chi_{1,l}^2$ $(l=1,\ldots,p)$ are i.i.d. $\chi_1^2$ random variables, and $c_1,\ldots,c_p$ are constants. As the first-order Wilks theorem is no longer available, there is no point of talking about the second-order Bartlett property. A general discussion on the first-order behavior of the EL ratio with plugged-in nuisance parameter estimators is available in Hjort et al. (2009).

### 5.3 Nonparametric imputation

For missing covariates, the imputation method can be employed as proposed in Wang and Chen (2009), based on a nonparametric kernel estimate of the conditional distribution of $X_i^{(1)}$ given $(X_i^{(2)}, Y_i)$. To simplify our notation, we write $(X_i^{(2)}, Y_i)$ as $Z_i$, which is $d_z =: (d_2+1)$-dimensional, and it is an always observable component of the data.

Let $F(x^{(1)}|Z_i)$ be the conditional distribution of $X_i^{(1)}$ given $(X_i^{(2)}, Y_i)$, and $W(\cdot)$ be a $d_z$-dimensional kernel function of the $q$th order satisfying

$$\int W(s_1,\ldots,s_{d_z})\,ds_1\cdots ds_{d_z} = 1,$$

$$\int s_i^l W(s_1,\ldots,s_{d_z})\,ds_1\cdots ds_{d_z} = 0 \quad \text{for any } i=1,\ldots,d_z \text{ and } 1\le l<q,$$

and $\int s_i^q W(s_1,\ldots,s_{d_z})ds_1\cdots ds_{d_z} \neq 0$. A kernel estimator of $F(x^{(1)}|Z_i)$ is

$$\hat{F}\big(x^{(1)}|Z_i\big) = \sum_{l=1}^{n}\frac{\delta_l W(\frac{Z_l-Z_i}{h})I(X_l^{(1)}\le x^{(1)})}{\sum_{l=1}^{n}\delta_l W(\frac{Z_l-Z_i}{h})}. \tag{45}$$

Here $h$ is the smoothing bandwidth, and $I(\cdot)$ is the $d_1$-dimensional indicator function. The property of the kernel estimator when there are no missing values is well understood in the literature, for instance, in Härdle (1990). Its properties in the context of missing values can be established in a standard fashion. For each missing $X_i^{(1)}$, we impute a missing $X_i^{(1)*}$ by randomly generating from the estimated conditional distribution $\hat{F}(x^{(1)}|Z_i)$. To control the variability due to the conditional distribution imputation, we make $\kappa$ independent draws $\{X_{i\nu}^{(1)*}\}_{\nu=1}^{\kappa}$ from $\hat{F}(x^{(1)}|Z_i)$. Specifically, let

$$\tilde{Z}_i(\beta) = \delta_i \frac{\partial m(X_i;\beta)}{\partial \beta}\{Y_i - m(X_i;\beta)\} + (1-\delta_i)\kappa^{-1}\sum_{l=1}^{\kappa}\frac{\partial m(X_{il}^{(1)*}, X_i^{(2)};\beta)}{\partial \beta}\{Y_i - m(X_{il}^{(1)*}, X_i^{(2)};\beta)\}$$

be the pseudo-estimating function for the regression parameters.

The EL for $\beta$ with the multiple imputed values for each missing $X_i^{(1)}$ is now

$$L_n(\beta) = \max \prod_{i=1}^{n} p_i$$

subject to $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n p_i \tilde{Z}_i(\beta) = 0$.

As shown in Wang and Chen (2009), Wilks' theorem is no longer valid for the EL ratio. Rather it is a weighted chi-square distribution similar to the case revealed in Wang and Rao (2002). A version of the bootstrap that reflects the missing value mechanism can be used to approximate the distribution of the EL ratio, which leads to likelihood-based confidence regions and hypothesis testing.

## 6 Regression with censored data

The EL method for censored data has a long history. It goes back to Thomas and Grunkemeier (1975), who proposed a method for constructing confidence intervals for survival probabilities when the data are subject to random right censoring, which directly motivates Owen's invention of the EL as recalled in Owen (2001). The EL method is in fact quite attractive for censored data, since its natural competitor, the normal method, often leads to complicated variance formulas caused by the censoring mechanism.

We focus here on the case of regression models where the response variable is subject to random right censoring. In this section we will try to summarize the many contributions that have been made in this context, making as before the distinction between parametric, nonparametric, and semiparametric models.

### 6.1 Parametric regression

Consider the accelerated failure time model $Y_i = \beta^T X_i + \varepsilon_i$, where $E(\varepsilon_i|X_i) = 0$, $\mathrm{Var}(\varepsilon_i|X_i) = \sigma^2$, $Y_i$ is the logarithm of the survival time, and $\beta$ is $p$-dimensional.

Instead of observing $Y_i$, we observe $T_i = \min(Y_i, C_i)$ and $\Delta_i = I(Y_i \leq C_i)$, where $C_i$ is a censoring variable, assumed to be independent of $Y_i$ given the $d = (p-1)$-dimensional vector $X_i$. The EL method for parametric regression described in Sect. 2 can be extended to censored data by replacing constraint (5), which is obtained from the normal equations for least squares estimators, by a similar equation for censored data. Many proposals exist in the literature for extending the least squares approach to censored data. See, e.g., Heuchenne and Van Keilegom (2007) for an overview of these proposals. Two popular approaches are the ones proposed by Buckley and James (1979) and Koul et al. (1981). In Qin and Jing (2001a) and Li and Wang (2003) the authors replace the normal equation (5) by the equation that lies on the basis of Koul et al. (1981)'s approach. More recently, Zhou and Li (2008) proposed an EL method, based on Buckley and James (1979)'s paper. In particular, for any vector $\beta$, let $e_i(\beta) = T_i - \beta^T X_i$ $(i = 1, \ldots, n)$, and for any distribution function $F$, whose support is given by the set of uncensored $e_i(\beta)$'s, define the empirical likelihood by

$$L_n(\beta, F) = \prod_{i=1}^{n} p_i^{\Delta_i} \left(1 - \sum_{e_j(\beta) \leq e_i(\beta)} p_j\right)^{1-\Delta_i},$$

where $p_i$ is the jump size of $F$ at $e_i(\beta)$. Note that for fixed $\beta$, this likelihood is maximized when $F$ equals the Kaplan–Meier estimator $\hat{F}_\beta$ based on $(e_i(\beta), \Delta_i)$ $(i = 1, \ldots, n)$. This motivates us to consider the following log EL ratio:

$$r_n(\beta_0) = -2 \log \frac{\sup_F L_n(\beta_0, F)}{L_n(\hat{\beta}, \hat{F}_{\hat{\beta}})},$$

where $\hat{\beta}$ is the Buckley–James estimator of $\beta_0$, and where the supremum in the numerator is taken over all distributions $F$ that satisfy the estimating equation of the Buckley–James estimator (see (4) in Zhou and Li 2008 for more details). An important feature of this EL ratio is that it is defined in terms of the likelihood for censored data, whereas other approaches (including Qin and Jing 2001a and Li and Wang 2003) use the complete-data likelihood and adjust the constraint under which the numerator is maximized for the presence of censoring.

It can now be shown that $r_n(\beta_0)$ converges in distribution to a $\chi^2_p$ random variable. Hence, this result can be used for doing inference for the vector $\beta_0$ without having to estimate the variance of the Buckley–James estimator, which is known to be quite cumbersome. It is easy to see that when no censoring is present, the denominator in the above expression equals $n^{-n}$, and the asymptotic result reduces to (13).

In survival analysis one often prefers to consider median regression, as opposed to mean regression, since survival data are often skewed to the right and the nonparametric estimation of the right tail of the error distribution is inaccurate when the data are subject to right censoring. Let us therefore consider the above linear regression model $Y_i = \beta^T X_i + \varepsilon_i$, but assume now that the conditional median of $\varepsilon_i$ given $X_i$ equals zero. In addition, assume that the censoring variable $C_i$ is independent of the vector of covariates $X_i$. For this model, Qin and Tsao (2003) considered the following

EL ratio, based on the likelihood for complete data:

$$R_n(\beta) = \max \prod_{i=1}^{n} (np_i)$$

subject to $p_i \geq 0$, $\sum_{i=1}^{n} p_i = 1$, and

$$\sum_{i=1}^{n} p_i X_i \left( \frac{I(T_i - \beta^T X_i \geq 0)}{1 - \hat{G}(\beta^T X_i)} - \frac{1}{2} \right) = 0,$$

where $\hat{G}$ is the Kaplan–Meier estimator of the censoring distribution $G$. This constraint is inspired by the normal-approximation-based method proposed by Ying et al. (1995). It can now be shown that $-2 \log R_n(\beta_0)$ converges in distribution to a weighted sum of independent $\chi_1^2$ variables. Note that the weights are caused by the estimator $\hat{G}$, whereas in the case of Zhou and Li (2008) the censoring distribution did not have to be estimated, since they work with the likelihood for censored data. Moreover, Zhou and Li (2008) do not have to make the rather restrictive assumption that $C_i$ is independent of $X_i$.

### 6.2 Nonparametric regression

We now focus on the case where the relation between the response $Y$ and a one-dimensional continuous covariate $X$ is completely unspecified (except for some smoothness assumptions), and the censoring variable $C$ is allowed to depend on $X$ in any (smooth) way. One is interested in doing inference for the conditional distribution $F(y|x) = P(Y \leq y | X = x)$.

Let $(X_i, T_i, \Delta_i)^T$ $(i = 1, \ldots, n)$ be an i.i.d. sample from the joint distribution of $(X, T, \Delta)$, where $T = \min(Y, C)$ and $\Delta = I(Y \leq C)$. Li and Van Keilegom (2002) considered the construction of EL confidence intervals for the survival probability $S(y|x) = 1 - F(y|x)$ for fixed $x$ and $y$. They also considered EL confidence bands when $y$ runs over an interval. Their method is based on localizing the censored data likelihood around the value $x$. In particular, we define the local log likelihood by

$$\begin{aligned}
&\log L_n\big(S(\cdot|x)\big) \\
&\quad = nh_n \sum_{i=1}^{n} W_i(x; h_n) \big\{ \Delta_i \log\big[ S(T_i - |x) - S(T_i|x) \big] + (1 - \Delta_i) \log S(T_i|x) \big\}, \\
&\quad = nh_n \sum_{i=1}^{n} W_i(x; h_n) \left\{ \Delta_i \log p_i + (1 - \Delta_i) \log \left( 1 - \sum_{T_j \leq T_i} p_j \right) \right\},
\end{aligned}$$

where

$$W_i(x; h_n) = \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)}$$

are Nadaraya–Watson weights (with kernel $K$ and bandwidth $h = h_n$), and $1 - S(\cdot|x)$ makes jumps of size $p_i$ at the uncensored $T_i$'s. In order to construct a confidence band for $S(y|x)$, we now define the EL ratio

$$R_n(p, t|x) = \frac{\sup\{L_n(S(\cdot|x)) : S(t|x) = p,\ S(\cdot|x) \in \Theta\}}{\sup\{L_n(S(\cdot|x)) : S(\cdot|x) \in \Theta\}},$$

where $\Theta$ is the space of all survival functions supported on $(0, \infty)$. Then, Li and Van Keilegom (2002) showed that for appropriate $0 < y_1 < y_2 < \infty$, the process

$$-2\frac{\hat{f}(x)}{\int K^2(u)\,du} \log R_n\big(S(y|x), y|x\big) \tag{46}$$

$(y_1 \le y \le y_2)$ converges weakly to the process $\{B^0(u)/\sqrt{u(1-u)}\}^2$, where $u = \sigma^2(y|x)/(1 + \sigma^2(y|x))$, $\sigma^2(\cdot|x)$ is the asymptotic variance of the cumulative hazard function of $Y$ given $X = x$, $\hat{f}(\cdot)$ is a kernel estimator of the density of $X$, and $B^0$ is a Brownian bridge on $[0, 1]$. Also note that for fixed $y$, the marginals of the process (46) converge to the marginals of the limiting process, which is a $\chi^2_1$ variable. Based on this result, it is now possible to construct confidence intervals and bands for the distribution $S(y|x)$ $(y_1 \le y \le y_2)$. For a similar result for the quantile function of $Y$ given $X$, we refer to Li and Van Keilegom (2002).

### 6.3 Semiparametric regression

An important semiparametric regression model in the context of survival data is without doubt the Cox proportional hazards model. The model is a special case of the so-called linear transformation model, given by

$$H(Y_i) = -\beta^T X_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{47}$$

where $H$ is an unknown monotone increasing (nuisance) function, $\beta$ a $p$-dimensional regression parameter vector, and $\varepsilon_i$ the error term with a known continuous distribution that is independent of the censoring variable $C_i$ and the covariate vector $X_i$. Let $\Lambda$ denote the cumulative hazard function of $\varepsilon_i$, i.e., $P(\varepsilon_i > t) = \exp\{-\Lambda(t)\}$. If $\Lambda(t) = \exp(t)$, then (47) becomes the proportional hazards model. On the other hand, if $\Lambda(t) = \log\{1 + \exp(t)\}$, then it becomes the proportional odds model. Let $(X_i^T, T_i, \Delta_i)^T$ $(i = 1, \ldots, n)$ be an i.i.d. sample coming from model (47). Lu and Liang (2006) showed how inference for the vector $\beta_0$ can be carried out using an EL approach. They base the empirical likelihood on the following martingale integral equation $(i = 1, \ldots, n)$:

$$E\left(\int_0^\infty X_i\big[dN_i(t) - Y_i(t)\,d\Lambda\big\{H(t) + \beta_0^T X_i\big\}\big]\right) = 0, \tag{48}$$

where $N_i(t) = \Delta_i I(T_i \le t)$ and $Y_i(t) = I(T_i \ge t)$. They showed that the log EL ratio associated with constraint (48), but with the unknown transformation $H$ replaced by an appropriate estimator, converges to a weighted sum of $p$ independent $\chi^2_1$ variables.

Other semiparametric models with censored data have been analyzed using EL methodology. See, e.g., Qin and Jing (2001b) and Wang and Li (2002) for the analysis of the partial linear model. The EL methodology for all these models can be seen as a special case of the general method developed by Hjort et al. (2009). For clarity of presentation, we do not explain their method in full generality, but we focus instead on a somewhat more narrow class of models, which is sufficiently large for the context of this paper. Consider a general semiparametric model depending on a response vector $Y$, a covariate vector $X$, a $p$-dimensional parameter vector $\beta$, and a nuisance function $g$. The true value of $\beta$ is denoted by $\beta_0$. The goal is to do inference for $\beta_0$ using an EL approach. Suppose that $\beta_0$ is the unique solution of the following system of equations in $\beta$:

$$E\big[m(X, Y, \beta, g)\big] = 0, \tag{49}$$

where $m$ is a $p$-dimensional function, and suppose that an estimator $\hat{g}$ of $g$ is available. For any $\beta$ and $g$, and for any i.i.d. sample $(X_i^T, Y_i^T)^T$ having the same distribution as $(X^T, Y^T)^T$, define the EL ratio $R_n(\beta, g)$ by

$$R_n(\beta, g) = \max \prod_{i=1}^{n} (np_i)$$

subject to $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$, and $\sum_{i=1}^n p_i m(X_i, Y_i, \beta, g) = 0$. Consider now the following four conditions:

$$P\big(R_n(\beta_0, \hat{g}) = 0\big) \to 0 \quad \text{as } n \to \infty;$$

$$n^{-1/2} \sum_{i=1}^{n} m(X_i, Y_i, \beta_0, \hat{g}) \xrightarrow{d} N(0, V_1);$$

$$n^{-1} \sum_{i=1}^{n} m(X_i, Y_i, \beta_0, \hat{g}) m^T(X_i, Y_i, \beta_0, \hat{g}) \xrightarrow{P} V_2;$$

$$\max_{i=1,\ldots,n} \big\| m(X_i, Y_i, \beta_0, \hat{g}) \big\| = o_P\big(n^{1/2}\big).$$

Under these conditions, the limiting distribution of $-2 \log R_n(\beta_0, \hat{g})$ is a weighted sum of $p$ independent $\chi_1^2$ variables, where the weights are the eigenvalues of $V_2^{-1} V_1$. When the estimation of these weights is cumbersome, Hjort et al. (2009) propose to approximate the limiting distribution by using a bootstrap approach, and they give generic conditions under which this bootstrap is consistent.

## 7 Goodness-of-fit tests

We have seen that the EL can be used to construct likelihood ratio confidence regions and hypothesis tests regarding regression parameters. In this section, we will show that EL is a natural device to formulate goodness-of-fit test statistics regarding the regression function $m(x) = E(Y_i | X_i = x)$, where $X_i$ is $d$-dimensional.

We start with testing for a parametric regression model

$$H_0 : m(\cdot) = m(\cdot; \beta_0) \quad \text{for a } \beta_0 \in B, \tag{50}$$

where $B$ is a compact set in $R^p$. Later we will extend it to tests for semiparametric models.

Naturally, goodness-of-fit tests can be constructed based on a distance between a nonparametric kernel regression estimator $\hat{m}(\cdot)$ and the parametric regression $m(\cdot; \hat{\beta})$, where $\hat{\beta}$ is an estimator of the finite-dimensional parameter $\beta$ under $H_0$. Under the null hypothesis $H_0$, this distance would take a smaller value than under the alternative $H_1$. For instance, the Härdle and Mammen (1993) test statistic is

$$T_{HM,n} = \left(nh^d\right)^{1/2} \int \left\{\hat{m}(x) - \tilde{m}(x; \hat{\beta})\right\}^2 \pi(x)\, dx,$$

where $\pi(\cdot)$ is a weight function, $\hat{m}$ is the kernel regression estimator (20) representing the model-free regression estimation, and

$$\tilde{m}(x; \hat{\beta}) = \frac{\sum_{i=1}^n K_h(x - X_i) m(X_i; \hat{\beta})}{\sum_{i=1}^n K_h(x - X_i)}$$

is a kernel-smoothed estimator of the parametric regression function under $H_0$. The purpose of applying the same kernel smoothing to the estimated parametric regression is to make the biases in the kernel estimation cancel each other. Asymptotic normality can be established for $T_{HM,n}$. Härdle and Mammen (1993) propose a wild bootstrap procedure to profile the finite-sample distribution of the test statistic.

An EL formulation for testing (50) consists of two steps. We first construct the EL for $m(x)$ at $\tilde{m}(x; \hat{\beta})$, which is

$$L_n\left\{\tilde{m}(x; \hat{\beta})\right\} = \max \prod_{i=1}^n p_i$$

subject to $\sum p_i = 1$ and $\sum p_i K_h(x - X_i)\{Y_i - \tilde{m}(x; \hat{\beta})\} = 0$. Let $r_n\{\tilde{m}(x; \hat{\beta})\} = -2\log[L_n\{\tilde{m}(x; \hat{\beta})\}n^n]$ be the log EL ratio. It may be seen by following similar steps to those outlined in Sect. 2 that

$$r_n\left\{\tilde{m}(x; \hat{\beta})\right\} = nh^d\left\{\hat{m}(x) - \tilde{m}(x; \hat{\beta})\right\}^2 V^{-1}(x)\left\{1 + o_p\left(h^{d/2}\right)\right\}, \tag{51}$$

where $V(x) = R(K)\sigma^2(x)/f(x)$, $f(\cdot)$ is the density of $X$, $R(K) = \int K^2(t)\, dt$, and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$. We then formulate the final test statistic

$$\mathcal{L}_n = \int r_n\left\{\tilde{m}(x; \hat{\beta})\right\} \pi(x)\, dx,$$

which has a leading order term

$$\int nh^d\left\{\hat{m}(x) - \tilde{m}(x; \hat{\beta})\right\}^2 V^{-1}(x)\pi(x)\, dx. \tag{52}$$

Hence, $\mathcal{L}_n$ is effectively a studentized $L_2$-distance between $\hat{m}(\cdot)$ and $\tilde{m}(\cdot;\hat{\beta})$. The EL formulation provides a studentization by $V^{-1}(x)$ automatically without having to estimate it explicitly. This is an attractive feature of the EL. In the current univariate regression situation, as shown in Chen et al. (2003a),

$$h^{-d/2}\{\mathcal{L}_n - \mu_0\} \xrightarrow{d} N\big(0, \sigma_0^2\big) \quad \text{as } n \to \infty,$$

where $\sigma_0^2 = 2K^{(4)}(0)\{K^{(2)}(0)\}^{-2}\int \pi^2(x)\,dx$ and $\mu_0 = 1 + h^{d/2}\int V^{-1}(x)\Delta_n^2(x) \times \pi(x)\,dx$. Here $\Delta_n(x)$ are uniformly bounded functions that define the difference between $m(x)$ and $m(x;\beta)$ in that $m(x) = m(x;\beta) + n^{-1/2}h^{-d/4}\Delta_n(x)$. Therefore, $\mathcal{L}_n$ is asymptotically pivotal under $H_0$.

The above EL formulation of the goodness-of-fit statistic can be extended to multiple regression curves with $Y_i$ being a $k$-variate response and $X_i$ still being a $d$-dimensional covariate. Let $m(x) = E(Y_i|X_i = x) = (m_1(x), \ldots, m_k(x))$ be the conditional mean consisting of $k$ regression curves on $R^d$ and $\Sigma(x) = \mathrm{Var}(Y_i|X_i = x)$ be a $k \times k$ matrix whose values may change along with the covariate. Let $m(\cdot) = m(\cdot, \beta, g) = (m_1(\cdot, \beta, g), \ldots, m_k(\cdot, \beta, g))$ be a working regression model, of which we would like to check its validity. Here, the form of $m$ is known up to a finite-dimensional parameter $\beta$ and an infinite-dimensional nuisance parameter $g$ where $\beta \in B \subset R^p$ and $g \in \mathcal{G}$ which is a complete metric space consisting of functions from $R^d$ to $R^q$ $(q \geq 1)$. This semiparametric regression model includes a wide range of parametric, semiparametric, and nonparametric regression models as special cases. In the absence of $g$, the model degenerates to a fully parametric model $m(\cdot) = m(\cdot, \beta)$, whereas the presence of $g$ covers a range of semiparametric models including the single or multiindex models and partially linear single-index models considered in Sect. 4. Nonparametric regression is also covered by taking the $\beta$-space as an empty set. The class also includes models with qualitative constraints, like additive models and models with shape constraints.

The goodness-of-fit hypotheses for the semiparametric regression are

$$H_0 : m(\cdot) = m(\cdot, \beta_0, g_0) \quad \text{for some } \beta_0 \in B \text{ and } g_0 \in \mathcal{G} \quad \text{versus}$$
$$H_1 : m(\cdot) \neq m(\cdot, \beta, g) \quad \text{for any } \beta \in B \text{ and any } g \in \mathcal{G}.$$

Let $\hat{\beta}$ be a $\sqrt{n}$-consistent estimator of $\beta_0$, and $\hat{g}$ be a consistent estimator of $g_0$ under a norm $\|\cdot\|_{\mathcal{G}}$ defined on the complete metric space $\mathcal{G}$. Any $\sqrt{n}$-consistent estimator of $\beta_0$ would be fine, for instance, the pseudo-likelihood estimator that assumes the residual distribution being normal. We suppose that $\hat{g}$ is a kernel estimator based on a kernel $L$ of order $s \geq 2$ and a bandwidth sequence $b$, most likely different from the bandwidth $h$ (defined below) used to estimate $m$. We will require that $\hat{g}$ converges to $g_0$ faster than $(nh^d)^{-1/2}$, the optimal rate in a completely $d$-dimensional nonparametric model. As demonstrated in Sect. 4, this can be easily satisfied since $g$ is of lower dimension than the saturated nonparametric model for $m$.

Again to cancel the bias due to kernel estimation for each $m_l(x)$, we smooth $m_l(x, \hat{\beta}, \hat{g})$ by the same kernel $K$ and bandwidth $h_l$ as in the kernel estimator $\hat{m}_l(x)$:

$$\tilde{m}_l(x, \hat{\beta}, \hat{g}) = \frac{\sum_{i=1}^n K_{h_l}(x - X_i)m_l(X_i, \hat{\beta}, \hat{g})}{\sum_{t=1}^n K_{h_l}(x - X_t)}$$

for $l = 1, \ldots, k$. Let $\tilde{m}(x, \hat{\beta}, \hat{g}) = (\tilde{m}_1(x, \hat{\beta}, \hat{g}), \ldots, \tilde{m}_k(x, \hat{\beta}, \hat{g}))^T$.

The EL formulation of the goodness-of-fit tests follows a similar line as the univariate parametric regression we have considered earlier in this section. We assume throughout that $h_l/h \to \beta_l$ as $n \to \infty$, where $h$ represents a baseline level of the smoothing bandwidth and $c_0 \le \min_l\{\beta_l\} \le \max_l\{\beta_l\} \le c_1$ for finite and positive constants $c_0$ and $c_1$ free of $n$.

Like our formulation for parametric regression shown above, we first conduct the empirical likelihood ratio for $m(x)$ evaluated at $\tilde{m}(x, \hat{\beta}, \hat{g})$ and then globalize by integrating the likelihood ratio to form the final test statistic.

Define at each fixed $x$,

$$\hat{Q}_i(x, \hat{\beta}) = \big(K_{h_1}(x - X_i)\big(Y_{i1} - \tilde{m}_1(x, \hat{\beta}, \hat{g})\big), \ldots, K_{h_k}(x - X_i)\big(Y_{ik} - \tilde{m}_k(x, \hat{\beta}, \hat{g})\big)\big)^T.$$

Let $\{p_i(x)\}_{i=1}^n$ be nonnegative empirical likelihood weights allocated to $\{(X_i, Y_i)\}_{i=1}^n$. The minus 2 log empirical likelihood ratio for the multiple conditional mean evaluated at $\tilde{m}(x, \hat{\beta}, \hat{g})$ is

$$r_n\big\{\tilde{m}(x, \hat{\beta}, \hat{g})\big\} = -2 \max \sum_{i=1}^n \log\big\{np_i(x)\big\}$$

subject to $p_i(x) \ge 0$, $\sum_{i=1}^n p_i(x) = 1$, and $\sum_{i=1}^n p_i(x)\hat{Q}_i(x, \hat{\beta}) = 0$. By introducing a vector of Lagrange multipliers $\lambda(x) \in R^k$, the optimal weights are given by

$$p_i(x) = \frac{1}{n}\big\{1 + \lambda^T(x)\hat{Q}_i(x, \hat{\beta})\big\}^{-1}, \tag{53}$$

where $\lambda(x)$ solves

$$\sum_{i=1}^n \frac{\hat{Q}_i(x, \hat{\beta})}{1 + \lambda^T(x)\hat{Q}_i(x, \hat{\beta})} = 0. \tag{54}$$

Integrating $r_n\{\tilde{m}(x, \hat{\beta}, \hat{g})\}$ over the weight function $\pi$ gives

$$\mathcal{L}_n = \int r_n\big\{\tilde{m}(x, \hat{\beta}, \hat{g})\big\}\pi(x)\,dx,$$

which is our EL test statistic based on the bandwidth vector $\mathbf{h} = (h_1, \ldots, h_k)^T$.

Define $\hat{\bar{Q}}(x, \hat{\beta}) = n^{-1}\sum_{i=1}^n \hat{Q}_i(x, \hat{\beta})$, $R(t) = \int K(u)K(tu)\,du$, and $V(x)$ is the product of $f(x)$ by a $k \times k$ matrix with $(j, l)$-element equal to $\beta_j^{-d} R(\beta_l/\beta_j)\sigma_{lj}(x)$. Note that $R(1) = R(K) =: \int K^2(u)\,du$ and that $\beta_j^{-d} R(\beta_l/\beta_j) = \beta_l^{-d} R(\beta_j/\beta_l)$ indicating that $V(x)$ is a symmetric matrix.

It may be shown that

$$\Lambda_n(\mathbf{h}) = nh^d \int \hat{\bar{Q}}^T(x, \beta_0)V^{-1}(x)\hat{\bar{Q}}(x, \beta_0)\pi(x)\,dx + o_p\big(h^{d/2}\big),$$

where $h^{d/2}$ is the stochastic order of the first term on the right-hand side if $d < 4r$. Here $r$ is the order of the kernel $K$. Since $\hat{\bar{Q}}(x, \beta_0) = f(x)\{\hat{m}(x) -$

$\tilde{m}(x, \beta_0, \hat{g})\}\{1 + o_p(1)\}$, $\hat{\tilde{Q}}(x, \beta_0)$ serves as a raw discrepancy measure between $\hat{m}(x) = (\hat{m}_1(x), \ldots, \hat{m}_k(x))$ and the hypothesized model $m(x, \beta_0, \hat{g})$. There is a key issue on how much each $\hat{m}_l(x) - \tilde{m}_l(x, \beta_0, \hat{g})$ contributes to the final statistic. The EL distributes the contributions according to $nh^d V^{-1}(x)$, the inverse of the covariance matrix of $\hat{\tilde{Q}}(x, \beta_0)$, which is the most natural choice. The nice thing about the EL formulation is that this is done without explicit estimation of $V(x)$ due to its internal standardization. Estimating $V(x)$ when $k$ is large can be challenging if not just tedious.

Let $(\gamma_{lj}(x))_{k\times k} = ((\beta_j^{-d} R(\beta_l/\beta_j)\sigma_{lj}(x))_{k\times k})^{-1}$,

$$\omega_{l_1,l_2,j_1,j_2}(\beta, K) = \iiint \beta_{l_2}^{-d} K(u)K(v)K\big\{(\beta_{j_2} z + \beta_{l_1} u)/\beta_{l_2}\big\} K(z + \beta_{j_1} v/\beta_{j_2})\, du\, dv\, dz,$$

$$\sigma^2(K, \Sigma) = 2 \sum_{l_1,l_2,j_1,j_2=1}^{k} \beta_{l_2}^{-d} \omega_{l_1,l_2,j_1,j_2}(\beta, K) \int \gamma_{l_1 j_1}(x)\gamma_{l_2 j_2}(x)\sigma_{l_1 l_2}(x)\sigma_{j_1 j_2}(x)\pi^2(x)\, dx,$$

which is a bounded quantity under certain assumptions given in Chen and Van Keilegom (2009). Chen and Van Keilegom (2009) establish the following asymptotic normality of $\mathcal{L}_n$ under $H_0$:

$$h^{-d/2}\{\mathcal{L}_n - k\} \xrightarrow{d} N\big(0, \sigma^2(K, \Sigma)\big) \quad \text{as } n \to \infty.$$

The convergence to the asymptotic normal distribution by the above two EL goodness-of-fit test statistics is quite slow since the test statistics are effectively $U$-statistics. This is the case for almost all goodness-of-fit statistics, EL or not. As a result, one tries to avoid carrying out the goodness-of-fit tests based on the asymptotic distribution. Rather, bootstrap resampling is used to better approximate the distributions of the test statistics. Chen and Van Keilegom (2009) outline a bootstrap algorithm for practical implementation.

## 8 Bibliographic notes

Owen (1988, 1990) are the two original papers that formally launched the empirical likelihood method. His work was motivated by the paper of Thomas and Grunkemeier (1975), who used a profile likelihood to construct confidence intervals for survival probabilities. Those authors showed that the confidence intervals have the desired property of respecting range, which is not generally held by normal-approximation-based methods. The idea of the empirical likelihood can be traced earlier, for instance, Hartley and Rao (1968), who applied the idea of the empirical likelihood in a survey sampling context. There were a series of papers on the general properties of the empirical likelihood method, which includes DiCiccio et al. (1989). Hall and La Scala (1990) gave the first review on the empirical likelihood. DiCiccio et al. (1991) showed

the Bartlett correction for parameters that are defined by smooth functions of means. A more general framework for empirical likelihood formulation than the framework of smooth functions of means is that of estimating equations, which includes parametric regression models as a special case. This framework allows the number of estimating equations to be larger than the number of parameters, which is a popular method in Econometrics, representing extra model information. Qin and Lawless (1994) established Wilks' theorem for the empirical likelihood in such context, and Chen and Cui (2006, 2007) showed that the Bartlett correction works.

The first paper that considered the empirical likelihood method for linear regression was Owen (1991). Chen (1993, 1994) established the Bartlett correction for linear regression. For generalized linear models, Kolaczyk (1994) formulated the EL based on the conditional mean aspect of the model; Chen and Cui (2003) considered adding extra constraints based on the conditional variance information within the GLIM to improve estimation efficiency.

Empirical likelihood for nonparametric regression was considered in Chen and Qin (2000) with a local linear kernel estimator and in Chen and Qin (2003) with the Nadaraya–Watson local constant kernel estimator. Both Wilks' theorem and Bartlett correction were established by carrying out undersmoothing to control the bias due to the kernel estimation.

In the context of semiparametric regression, Shi and Lau (2000) considered a partially linear regression with fixed design and obtained a similar result as Wang and Jing (2003) did for random design. They considered general weight functions satisfying certain regularity conditions. Lu (2009) considered the extension of Wang and Jing (2003)'s paper to the context of heteroscedastic regression. Hu et al. (2009) applied the empirical likelihood methodology to varying-coefficient partially linear errors-in-variables models, and Liang and Qin (2008) showed Wilks' theorem when the covariate $X$ is missing with probability depending on the response $Y$ and the covariate $Z$ (whose effect on $Y$ is modeled nonparametrically).

Using empirical likelihood for inference on the mean of the response variable when the response is subject to missingness at random was considered in Wang and Rao (2002) for a nonparametric regression model and Wang et al. (2004) for a semiparametric partially linear regression model. See also Wang and Veraverbeke (2006) for an approach based on auxiliary information. Chen et al. (2003b, 2008) considered inference when there are surrogates for the missing values. Qin and Zhang (2007) considered missing responses in the context of observational studies. Wang and Chen (2009) proposed the multiple nonparametric imputation for general estimating equations where the missing values can be either in the response or the covariates.

The literature on the EL methodology for censored data is becoming very extensive. For parametric mean regression, Zhou et al. (2006) proposed a generalized linear model for modeling health care costs and studied an EL procedure for this model. Zhao and Wang (2008) applied an EL approach to do inference for quality-adjusted lifetime data. For parametric median regression, we cite Whang (2006), who used a smoothed EL approach to obtain better performance in practice than the classical EL method. See also Zhao and Chen (2008). Zhou (2005) used an empirical likelihood analysis of a rank estimator in the accelerated failure time model, whereas Zhou (1992) proposed an $M$-estimation procedure. The EL methodology for the Cox model has been first considered by Qin and Jing (2001c).

Using EL to test for goodness-of-fit of parametric time series regression was considered in Chen et al. (2003a). Fan and Zhang (2004) propose a sieve EL test for testing a varying-coefficient regression model that extends the generalized likelihood ratio test of Fan et al. (2001). They demonstrate that "Wilks' phenomenon" continues to hold under general error distributions. Tripathi and Kitamura (2003) propose an EL test for conditional moment restrictions. For testing semiparametric regression models, Chen and Van Keilegom (2009) developed a smoothing-based EL approach, whereas Van Keilegom et al. (2008) use an EL approach based on marked empirical processes.

**Acknowledgements** The first author's research is supported by National Science Foundation grants 0604563 and 0714978, and a grant from Guanghua School of Management, Peking University. The second author acknowledges financial support from IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy), and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement No. 203650. Part of the review reported in this paper was carried out while the second author was visiting Guanghua School of Management, Peking University.

## References

Barndorff-Nielsen OE, Cox DR (1984) Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. J R Stat Soc Ser B 46:483–495

Barndorff-Nielsen OE, Hall PG (1988) On the level-error after Bartlett adjustment of the likelihood ratio statistic. Biometrika 75:374–378

Bartlett MS (1937) Properties of sufficiency and statistical tests. Proc R Soc A 160:268–282

Buckley JJ, James IR (1979) Linear regression with censored data. Biometrika 66:429–436

Chen SX (1993) On the accuracy of empirical likelihood confidence regions for linear regression model. Ann Inst Stat Math 45:621–637

Chen SX (1994) Empirical likelihood confidence intervals for linear regression coefficients. J Multivar Anal 49:24–40

Chen SX (1996) Empirical likelihood confidence intervals for nonparametric density estimation. Biometrika 83:329–341

Chen SX, Cui H (2003) An extended empirical likelihood for generalized linear models. Stat Sin 13:69–81

Chen SX, Cui H-J (2006) On Bartlett correction of empirical likelihood in the presence of nuisance parameters. Biometrika 93:215–220

Chen SX, Cui H-J (2007) On the second order properties of empirical likelihood with moment restrictions. J Econom 141:492–516

Chen SX, Qin Y-S (2000) Empirical likelihood confidence interval for a local linear smoother. Biometrika 87:946–953

Chen SX, Qin Y-S (2003) Coverage accuracy of confidence intervals in nonparametric regression. Acta Math Appl Sin Engl Ser 19:387–396

Chen SX, Van Keilegom I (2009) A goodness-of-fit test for parametric and semiparametric models in multiresponse regression. Bernoulli (to appear)

Chen SX, Härdle W, Li M (2003a) An empirical likelihood goodness-of-fit test for time series. J R Stat Soc, Ser B 65:663–678

Chen SX, Leung DHY, Qin J (2003b) Information recovery in a study with surrogate endpoints. J Am Stat Assoc 98:1052–1062

Chen SX, Leung DHY, Qin J (2008) Improved semiparametric estimation using surrogate data. J R Stat Soc, Ser B 70:803–823

DiCiccio T, Hall P, Romano J (1989) Comparison of parametric and empirical likelihood functions. Biometrika 76:465–476

DiCiccio T, Hall P, Romano J (1991) Empirical likelihood is Bartlett-correctable. Ann Stat 19:1053–1061

Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman & Hall, London

Fan J, Zhang J (2004) Sieve empirical likelihood ratio tests for nonparametric functions. Ann Stat 32:1858–1907
Fan J, Zhang C, Zhang J (2001) Generalized likelihood ratio statistics and Wilks phenomenon. Ann Stat 29:153–193
Hall P (1991) Edgeworth expansions for nonparametric density estimators, with applications. Statistics 22:215–232
Hall P, La Scala B (1990) Methodology and algorithms of empirical likelihood. Int Stat Rev 58:109–127
Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50:1029–1054
Härdle W (1990) Applied Nonparametric Regression. Cambridge University Press, Cambridge
Härdle W, Mammen E (1993) Comparing nonparametric versus parametric regression fits. Ann Stat 21:1926–1947
Hartley HO, Rao JNK (1968) A new estimation theory for sample surveys. Biometrika 55:547–557
Hayakawa T (1977) The likelihood ratio criterion and the asymptotic expansion of its distribution. Ann Inst Stat Math 29:359–378
Heuchenne C, Van Keilegom I (2007) Polynomial regression with censored data based on preliminary nonparametric estimation. Ann Inst Stat Math 59:273–298
Hjort NL, McKeague IW, Van Keilegom I (2009) Extending the scope of empirical likelihood. Ann Stat 37:1079–1115
Hu X, Wang Z, Zhao Z (2009) Empirical likelihood for semiparametric varying-coefficient partially linear errors-in-variables models. Stat Prob Lett 79:1044–1052
Kolaczyk ED (1994) Empirical likelihood for generalized linear model. Stat Sin 4:199–218
Koul H, Susarla V, Van Ryzin J (1981) Regression analysis with randomly right-censored data. Ann Stat 9:1276–1288
Lawley DN (1956) A general method for approximating the distribution of likelihood ratio criteria. Biometrika 43:295–303
Li G, Van Keilegom I (2002) Likelihood ratio confidence bands in nonparametric regression with censored data. Scand J Stat 29:547–562
Li G, Wang Q-H (2003) Empirical likelihood regression analysis for right censored data. Stat Sin 13:51–68
Liang H, Qin Y (2008) Empirical likelihood-based inferences for partially linear models with missing covariates. Aust N Z J Stat 50:347–359
Lu W, Liang Y (2006) Empirical likelihood inference for linear transformation models. J Multivar Anal 97:1586–1599
Lu X (2009) Empirical likelihood for heteroscedastic partially linear models. J Multivar Anal 100:387–396
McCullagh P, Nelder JA (1983) Generalized linear models. Chapman & Hall, London
Molanes-López E, Van Keilegom I, Veraverbeke N (2009) Empirical likelihood for non-smooth criterion functions. Scand J Stat 36:413–432
Owen AB (1988) Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75:237–249
Owen AB (1990) Empirical likelihood confidence regions. Ann Stat 18:90–120
Owen AB (1991) Empirical likelihood for linear models. Ann Stat 19:1725–1747
Owen A (2001) Empirical likelihood. Chapman & Hall, New York
Tripathi G, Kitamura Y (2003) Testing conditional moment restrictions. Ann Stat 31:2059–2095
Qin G, Jing B-Y (2001a) Empirical likelihood for censored linear regression. Scand J Stat 28:661–673
Qin G, Jing B-Y (2001b) Censored partial linear models and empirical likelihood. J Multivar Anal 78:37–61
Qin G, Jing B-Y (2001c) Empirical likelihood for Cox regression model under random censoring. Commun Stat Simul Comput 30:79–90
Qin G, Tsao M (2003) Empirical likelihood inference for median regression models for censored survival data. J Multivar Anal 85:416–430
Qin J, Lawless J (1994) Empirical likelihood and general estimating equations. Ann Stat 22:300–325
Qin J, Zhang B (2007) Empirical-likelihood-based inference in missing response problems and its application in observational studies. J R Stat Soc, Ser B 69:101–122
Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55
Rubin DB (1976) Inference and missing data. With comments by RJA Little and a reply by the author. Biometrika 63:581–592

Shi J, Lau T-S (2000) Empirical likelihood for partially linear models. J Multivar Anal 72:132–148
Thomas DR, Grunkemeier GL (1975) Confidence interval estimation of survival probabilities for censored data. J Am Stat Assoc 70:865–871
Tsao M (2004) Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. Ann Stat 32:1215–1221
Van Keilegom I, Sánchez Sellero C, González Manteiga W (2008) Empirical likelihood based testing for regression. Electr J Stat 2:581–604
Wang D, Chen SX (2009) Empirical likelihood for estimating equation with missing values. Ann Stat 37:490–517
Wang L, Veraverbeke N (2006) Empirical likelihood in a semi-parametric model for missing response data. Commun Stat Theory Methods 35:625–639
Wang Q-H, Jing B-Y (2003) Empirical likelihood for partial linear models. Ann Inst Stat Meth 55:585–595
Wang Q-H, Li G (2002) Empirical likelihood semiparametric regression analysis under random censorship. J Multivar Anal 83:469–486
Wang Q, Rao JNK (2002) Empirical likelihood-based inference under imputation for missing response data. Dedicated to the memory of Lucien Le Cam. Ann Stat 30:896–924
Wang Q, Linton O, Härdle W (2004) Semiparametric regression analysis with missing response at random. J Am Stat Assoc 99:334–345
Whang Y-J (2006) Smoothed empirical likelihood methods for quantile regression models. Econom Theory 22:173–205
Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9:60–62
Xue L-G, Zhu L (2006) Empirical likelihood for single-index models. J Multivar Anal 97:1295–1312
Ying Z, Jung SH, Wei LJ (1995) Survival analysis with median regression models. J Am Stat Assoc 90:178–184
Zhao Y, Chen F (2008) Empirical likelihood inference for censored median regression model via nonparametric kernel estimation. J Multivar Anal 99:215–231
Zhao Y, Wang H (2008) Empirical likelihood inference for the regression model of mean quality-adjusted lifetime with censored data. Can J Stat 36:463–478
Zhou M (1992) M-estimation in censored linear models. Biometrika 79:837–841
Zhou M (2005) Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model. Biometrika 92:492–498
Zhou M, Li G (2008) Empirical likelihood analysis of the Buckley–James estimator. J Multivar Anal 99:649–664
Zhou XH, Qin GS, Lin HZ, Li G (2006) Inferences in censored cost regression models with empirical likelihood. Stat Sin 16:1213–1232
Zhu L, Xue L (2006) Empirical likelihood confidence regions in a partially linear single-index model. J R Stat Soc, Ser B 68:549–570