应用概率统计 第 38 卷
第 2 期 2022 年 4 月

Chinese Journal of Applied Probability and Statistics
Apr., 2022, Vol. 38, No. 2, pp. 303-315
doi: 10.3969/j.issn.1001-4268.2022.02.009

Uncertain Semiparametric Model with Bernstein Polynomials *

DING Jianhua[1,2⋆] ZHANG Hongyu[1] ZHANG Zhiqiang[1]
([1] *Department of Statistics, Shanxi Datong University, Datong, 037009, China*)
([2] *KLATASDS-MOE, Shanghai, 200062, China*)

Abstract: Assuming the observations are imprecise and modeling the observations by uncertain variables, this paper proposes statistical inferences for uncertain semiparametric regression model when nonparametric function is subject to monotonicity constraint. Monotonic Bernstein polynomials are used to approximate the nonparametric function and quadratic programming algorithm is used to compute the estimate. A numerical example is given to illustrate the proposed methods.

Keywords: uncertain variable; uncertain semiparametric regression; Bernstein polynomials; cross validation method

2020 Mathematics Subject Classification: 62G08

Citation: DING J H, ZHANG H Y, ZHANG Z Q. Uncertain semiparametric model with Bernstein polynomials [J]. Chinese J Appl Probab Statist, 2022, 38(2): 303–315.

§1. Introduction

When the samples of response variable and (or) explanatory variables are imprecise, the uncertain variables can be employed to model the imprecise observations. The use for the uncertain variables was developed by many authors, such as Yao [1], Wen et al. [2], Wang et al. [3], Zhao et al. [4], Lio and Liu [5]. Nowadays, uncertain variables have been applied to many fields including uncertain risk analysis [6], uncertain programming [7] and uncertain process [8], etc. Uncertain statistics is a method to collect and interpret expert's data. Liu [9] proposed the least squares estimation of the unknown parameters in the uncertainty distribution. To estimate the unknown parameters in the uncertain regression models, the principle of least squares was suggested by Yao and Liu [10]. Lio and Liu [11] provided the residual analysis of uncertain linear regression models. Lio and Liu [12] proposed

---

*The project was supported by the National Statistical Science Research Project (Grant No. 2020LY010), the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, ECNU and the Fundamental Research Funds for the Central Universities.

⋆Corresponding author, E-mail: yp_dingjianhua@163.com.

Received April 29, 2021. Revised May 16, 2021.

maximum likelihood estimation of uncertain linear regression models. Song and Fu[13] propose uncertain multivariable linear regression model which has multiple response variables. Fang and Hong[14] applied the logarithmic, square root or reciprocal transformation to alleviate possible nonlinearity problems and estimate the disturbance terms for the uncertain nonlinear parametric models. Yang and Ni[15] proposed least squares estimation for uncertain moving average model. Liu and Yang[16] proposed least absolute estimate of uncertain parametric regression model. Liu and Jia[17] proposed cross-validation for uncertain Chapman-Richards model. Ding and Zhang[18] proposed B-Splines estimate and local polynomials estimate of uncertain nonparametric regression model.

In this paper, we study the uncertain semiparametric regression when the nonparametric function satisfies monotonicity. The rest of this paper is organized as follows. In Section 2, we introduce some basic results about uncertain variables. The statistical inferences for uncertain semiparametric regression model is proposed in Section 3. A numerical example is given to illustrate the proposed methods in Section 4. Some conclusions are made in Section 5.

§2. Preliminary

In this section, some basic concepts and formulas in uncertainty theory are introduced as follows.

Definition 1[19] Let Γ be a nonempty set, and let $\mathscr{L}$ be a σ-algebra over Γ. An uncertain measure is a function $\mathscr{M}:\mathscr{L}\to[0,1]$ such that

Axiom 1. (Normality Axiom) $\mathscr{M}\{\Gamma\}=1$ for the universal set Γ.

Axiom 2. (Duality Axiom) $\mathscr{M}\{\Lambda\}+\mathscr{M}\{\Lambda^c\}=1$ for any event Λ.

Axiom 3. (Subadditivity Axiom) For every countable sequence of events $\{\Lambda_i\}$, $i=1,2,\cdots$, we have

$$\mathscr{M}\Big\{\bigcup_{i=1}^{\infty}\Lambda_i\Big\}\leqslant\sum_{i=1}^{\infty}\mathscr{M}\{\Lambda_i\}.$$

A set $\Lambda\in\mathscr{L}$ is called an event. The uncertain measure $\mathscr{M}\{\Lambda\}$ indicates the degree of belief that Λ will occur. The triplet $(\Gamma,\mathscr{L},\mathscr{M})$ is called an uncertainty space. In order to obtain an uncertain measure of compound event, a product uncertain measure was defined by the following fourth axiom.

Axiom 4. (Product Axiom)[20] Let $(\Gamma_k,\mathscr{L}_k,\mathscr{M}_k)$ be uncertainty spaces for $k=1,2,\cdots$. The product uncertain measure $\mathscr{M}$ is an uncertain measure on the product σ-algebra

$\mathscr{L}_1 \times \mathscr{L}_2 \times \cdots$ satisfying

$$\mathscr{M}\Big\{\prod_{k=1}^{\infty} \Lambda_k\Big\} = \bigwedge_{k=1}^{\infty} \mathscr{M}_k\{\Lambda_k\}, \tag{1}$$

where Λ_k are arbitrarily chosen events from $\mathscr{L}_k$ for $k = 1, 2, \cdots$, respectively.

Definition 2 [19] An uncertain variable is a measurable function from an uncertainty space $(\Gamma, \mathscr{L}, \mathscr{M})$ to the set of real numbers, i.e., $\{\xi \in B\}$ is an event for any Borel set B.

Uncertain variable is introduced to model the quantity with human uncertainty.

Definition 3 [19] The uncertainty distribution Φ of an uncertain variable ξ is defined by $\Phi(x) = \mathscr{M}\{\xi \leqslant x\}$, for any real number x.

An uncertainty distribution Φ is said to be regular if it is a continuous and strictly increasing function with respect to x at which $0 < \Phi(x) < 1$, and

$$\lim_{x \to -\infty} \Phi(x) = 0, \qquad \lim_{x \to \infty} \Phi(x) = 1.$$

If ξ is an uncertain variable with regular uncertainty distribution $\Phi(x)$, the inverse function $\Phi^{-1}(\alpha)$ is called the inverse uncertainty distribution of ξ [9].

Definition 4 [20] The uncertain variables $\xi_1, \xi_2, \cdots, \xi_m$ are said to be independent if

$$\mathscr{M}\Big\{\bigcap_{i=1}^{m} \{\xi_i \in B_i\}\Big\} = \bigwedge_{i=1}^{m} \mathscr{M}\{\xi_i \in B_i\},$$

for any Borel sets $B_1, B_2, \cdots, B_m$ of real numbers.

Theorem 5 [9] Let $\xi_1, \xi_2, \cdots, \xi_n$ be independent uncertain variables with regular uncertainty distributions $\Phi_1, \Phi_2, \cdots, \Phi_n$, respectively. If f is strictly increasing with respect to $\xi_1, \xi_2, \cdots, \xi_m$ and strictly decreasing with respect to $\xi_{m+1}, \xi_{m+2}, \cdots, \xi_n$, then $\xi = f(\xi_1, \xi_2, \cdots, \xi_n)$ is an uncertain variable with uncertainty distribution

$$\Phi(x) = \sup_{f(x_1, x_2, \cdots, x_n) < x} \Big\{ \min_{1 \leqslant i \leqslant m} \Phi_i(x_i) \wedge \min_{m+1 \leqslant i \leqslant n} [1 - \Phi_i(x_i)] \Big\},$$

and with inverse uncertainty distribution

$$\Psi^{-1}(\alpha) = f(\Phi_1^{-1}(\alpha), \cdots, \Phi_m^{-1}(\alpha), \Phi_{m+1}^{-1}(1-\alpha), \cdots, \Phi_n^{-1}(1-\alpha)). \tag{2}$$

Definition 6 [19] The expected value of an uncertain variable ξ is defined by

$$\mathsf{E}[\xi] = \int_0^{\infty} \mathscr{M}\{\xi \geqslant x\} \mathrm{d}x - \int_{-\infty}^{0} \mathscr{M}\{\xi \leqslant x\} \mathrm{d}x,$$

provided that at least one of the two integrals is finite.

Assume that the uncertain variable ξ has an uncertainty distribution Φ. If the expected value $\mathsf{E}[\xi]$ exists, then

$$\mathsf{E}[\xi]=\int_0^{\infty}[1-\Phi(x)]\mathrm{d}x-\int_{-\infty}^{0}\Phi(x)\mathrm{d}x.$$

Example: An uncertain variable ξ is called linear if it has a linear uncertainty distribution

$$\Phi(x)=\begin{cases}0, & \text{if } x\leqslant a;\\ \dfrac{x-a}{b-a}, & \text{if } a\leqslant x\leqslant b;\\ 1, & \text{if } y\geqslant b,\end{cases}\tag{3}$$

denoted by $\mathscr{L}(a,b)$, where a and b are real numbers with $a<b$. The linear uncertain variable $y\sim\mathscr{L}(a,b)$ has an expected value $\mathsf{E}[y]=(a+b)/2$. The inverse uncertainty distribution of linear uncertain variable $\mathscr{L}(a,b)$ is $\Phi^{-1}(\alpha)=(1-\alpha)a+b\alpha$.

Definition 7 [1] Let ξ be an uncertain variable with uncertainty distribution Φ and finite expected value $\mathsf{E}[\xi]$. Then

$$V[\xi]=\mathsf{E}[(\xi-\mathsf{E}[\xi])^2]=\int_{-\infty}^{+\infty}(y-\mathsf{E}[\xi])^2\mathrm{d}\Phi(y).$$

Let ξ be an uncertain variable with an uncertainty distribution Φ. If its kth moment $\mathsf{E}[\xi^k]$ exists, then

$$\mathsf{E}[\xi^k]=\int_{-\infty}^{\infty}x^k\mathrm{d}\Phi(x).$$

Furthermore, if Φ is regular, then

$$\mathsf{E}[\xi]=\int_0^1\Phi^{-1}(\alpha)\mathrm{d}\alpha,\qquad \mathsf{E}[\xi^2]=\int_0^1[\Phi^{-1}(\alpha)]^2\mathrm{d}\alpha,\tag{4}$$

$$V[\xi]=\int_0^1[\Phi^{-1}(\alpha)-\mathsf{E}[\xi]\,]^2\mathrm{d}\alpha.\tag{5}$$

Let ξ be an uncertain variable with uncertainty distribution Φ, and $f(x)$ be a strictly increasing function, by Theorem 5, we have

$$\mathsf{E}[f(\xi)]=\int_0^1 f[\Phi^{-1}(a)]\mathrm{d}a.\tag{6}$$

§3. Statistical Inferences for Uncertain Semiparametric Model

In this section, we propose statistical inferences including estimation, residual analysis and forecast value for uncertain semiparametric model. Let $(x_1,x_2,\cdots,x_p,z)$ be a vector of explanatory variables, and let y be a response variable. Assume the relationship between $(x_1, x_2, \cdots, x_p, z)$ and y can be expressed by the semiparametric model:

$$y = \sum_{j=1}^{p} \beta_j x_j + f(z) + \epsilon, \tag{7}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_n)$ is a vector of unknown parameters, $f(\cdot)$ is an unknown nonparametric function satisfying monotonicity, ϵ is a disturbance term. $\sum\limits_{j=1}^{p} \beta_j x_j$ is called parametric part, $f(z)$ is called nonparametric part. Suppose that there are a set of imprecisely observed data,

$$(\widetilde{x}_{i1}, \widetilde{x}_{i2}, \cdots, \widetilde{x}_{ip}, \widetilde{z}_i, \widetilde{y}_i), \qquad i = 1, 2, \cdots, n,$$

where $\widetilde{x}_{i1}, \widetilde{x}_{i2}, \cdots, \widetilde{x}_{ip}, \widetilde{z}_i, \widetilde{y}_i$ are uncertain variables with uncertainty distributions $\Gamma_{i1}, \Gamma_{i2}, \cdots, \Gamma_{ip}, \Phi_i, \Psi_i$, $i = 1, 2, \cdots, n$, respectively. Based on the imprecisely observed data, we suggest the least squares estimates of $\boldsymbol{\beta}$ and f are the solution of the following minimization problem,

$$\min_{\boldsymbol{\beta}, f \in \mathscr{M}} \sum_{i=1}^{n} \mathsf{E}\Big[\Big(\widetilde{y}_i - \sum_{j=1}^{p} \beta_j \widetilde{x}_{ij} - f(\widetilde{z}_i)\Big)^2\Big], \tag{8}$$

where $\mathscr{M} = \{f : f \text{ is strictly monotone function}\}$.

1) Bernstein Estimation

Since the change of variables specified by $t = (z - a)/(b - a)$ maps $z \in [a, b]$ to $t \in [0, 1]$ without changing the max norm of any function, we can restrict our attention to continuous functions $f(z)$ on $z \in [0, 1]$ without loss of generality. Bernstein polynomials allow a great deal of flexibility in modeling the shape of the relationship between variables. For a continuous function such as $f(z)$ on $[0, 1]$, the approximating Bernstein polynomial of order m is given by

$$f(z) \approx B(z; f) = \sum_{j=0}^{m} f\Big(\frac{j}{m}\Big)\binom{m}{j} z^j (1 - z)^{m-j} = \sum_{j=0}^{m} b_j \delta_j(z),$$

where $\delta_j(z) = \binom{m}{j} z^j (1 - z)^{m-j}$ are the Bernstein basis polynomials, $b_j = f(j/m)$ are corresponding coefficients, $j = 0, 1, \cdots, m$. By the Weierstrass theorem, $B(\cdot; f) \to f(\cdot)$ uniformly over $[0, 1]$ as $m \to \infty$ [21]. Since the first derivative of $B(z; f)$ can be written as

$$B'(z; f) = m \sum_{j=0}^{m-1} (b_{j+1} - 2b_j)\binom{m-1}{j} z^j (1 - z)^{m-1-j},$$

$B(z;f)$ is strictly increasing function if $b_0 \leqslant b_1 \leqslant \cdots \leqslant b_m$. The increasing monotonicity is expressed in matrix form as

$$\boldsymbol{A}\boldsymbol{b} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots \\ & & \ddots & & \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix} \geqslant \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where $\boldsymbol{A}$ is $m \times (m+1)$ constrainted matrix, $\boldsymbol{b} = (b_0, b_1, \cdots, b_m)$. Similarly, $B(z;f)$ is strictly decreasing function if $b_0 \geqslant b_1 \geqslant \cdots \geqslant b_m$, and expressed in matrix form as $\boldsymbol{A}\boldsymbol{b} \leqslant \boldsymbol{0}$. For a vector $\boldsymbol{c}$, we use $\boldsymbol{c} \leqslant \boldsymbol{0}$ to denote the fact that the inequality is satisfied componentwise.

Theorem 8 Suppose that $\widetilde{x}_{i1}, \widetilde{x}_{i2}, \cdots, \widetilde{x}_{ip}, \widetilde{z}_i, \widetilde{y}_i$, $i = 1, 2, \cdots, n$, are a set of imprecisely observations, where $\widetilde{x}_{i1}, \widetilde{x}_{i2}, \cdots, \widetilde{x}_{ip}, \widetilde{z}_i, \widetilde{y}_i$ are independent uncertain variables with regular uncertainty distributions $\Gamma_{i1}, \Gamma_{i2}, \cdots, \Gamma_{ip}, \Phi_i, \Psi_i$, respectively. Then the least squares estimate of $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ and $f(z)$ in the semiparametric regression model (7) is the optimal solution of the following problem:

$$\min_{\boldsymbol{\beta}, \boldsymbol{b}} \sum_{i=1}^{n} \int_0^1 \Big[\Psi_i^{-1}(\alpha) - \sum_{j=1}^{p} \beta_j \Pi_{ij}^{-1}(\alpha, \beta_j) - \sum_{j=0}^{m} b_j \cdot \delta_j(\Upsilon_i^{-1}(\alpha, f))\Big]^2 \mathrm{d}\alpha, \tag{9}$$

with respcet to $\boldsymbol{\beta}$, $\boldsymbol{b}$ subject to $\boldsymbol{A}\boldsymbol{b} \geqslant \boldsymbol{0}$ (or $\boldsymbol{A}\boldsymbol{b} \leqslant \boldsymbol{0}$), where

$$\Pi_{ij}^{-1}(\alpha, \beta_j) = \begin{cases} \Gamma_{ij}^{-1}(1-\alpha), & \text{if } \beta_j > 0; \\ \Gamma_{ij}^{-1}(\alpha), & \text{if } \beta_j < 0, \end{cases}$$

and

$$\Upsilon_i^{-1}(\alpha, f) = \begin{cases} \Phi_i^{-1}(1-\alpha), & \text{if } f \text{ is strictly increasing;} \\ \Phi_i^{-1}(\alpha), & \text{if } f \text{ is strictly decreasing,} \end{cases}$$

for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$.

Proof By (8), the least squares estimate of $\boldsymbol{\beta}$ and f is the optimal solution of the minimization problem,

$$\min_{\boldsymbol{\beta}, f \in \mathscr{M}} \sum_{i=1}^{n} \mathsf{E}\Big[\Big(\widetilde{y}_i - \sum_{j=1}^{p} \beta_j \widetilde{x}_{ij} - f(\widetilde{z}_i)\Big)^2\Big].$$

For each i, it follows from [22] that the inverse uncertainty distribution of $\widetilde{y}_i - \sum\limits_{j=1}^{p} \beta_j \widetilde{x}_{ij} - f(\widetilde{z}_i)$ is

$$F_i^{-1}(\alpha) = \Psi_i^{-1}(\alpha) - \sum_{j=1}^{p} \beta_j \Pi_{ij}^{-1}(\alpha, \beta_j) - f(\Upsilon_i^{-1}(\alpha, f)).$$

By (4), we can get

$$\begin{aligned}&\mathsf{E}\Big[\Big(\widetilde{y}_i-\sum_{j=1}^{p}\beta_j\widetilde{x}_{ij}-f(\widetilde{z}_i)\Big)^2\Big]\\&=\int_0^1\Big[\Psi_i^{-1}(\alpha)-\sum_{j=1}^{p}\beta_j\Pi_{ij}^{-1}(\alpha,\beta_j)-f(\Upsilon_i^{-1}(\alpha,f))\Big]^2\mathrm{d}\alpha.\end{aligned}$$

Substituting $f(\Upsilon_i^{-1}(\alpha,f))$ with $\sum\limits_{j=0}^{m}b_j\cdot\delta_j(\Upsilon_i^{-1}(\alpha,f))$, then the minimization problem (8) is equivalent to the minimization problem (9). The theorem is verified. □

The above optimization problem can be effectively solved by the general quadratic programming[23]. Quadratic programming has also been used to solve the linear inequality constraints such as $\boldsymbol{Ab}\geqslant\mathbf{0}$. In this study, we use the available R package *quadprog* by Turlach and Weingessel[24] to solve quadratic programming problem. The order m may be chosen by using the ideas of cross-validation method, see Section 4 for more details.

2) Residual Analysis and Forecast Value

Definition 9 Let $(\widetilde{x}_{i1},\widetilde{x}_{i2},\cdots,\widetilde{x}_{ip},\widetilde{z}_i,\widetilde{y}_i)$, $i=1,2,\cdots,n$, be a set of imprecisely observations, and suppose the fitted regression model is

$$y_i=\sum_{j=1}^{p}\widehat{\beta}_jx_{ij}+\widehat{f}(z_i).$$

Then for each i, the term

$$\widehat{\epsilon}_i=\widetilde{y}_i-\sum_{j=1}^{p}\widehat{\beta}_j\widetilde{x}_{ij}-\widehat{f}(\widetilde{z}_i)$$

is called the i-th residual, for $i=1,2,\cdots,n$.

Now assume that the disturbance term ϵ is an uncertain variable with $\mathsf{E}[\epsilon]=e$ and $V[\epsilon]=\sigma^2$. Then we use the average of the expected values of residuals,

$$\widehat{e}=\frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[\widehat{\epsilon}_i] \tag{10}$$

to estimate the expected value of the disturbance term ϵ, and

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[(\widehat{\epsilon}_i-\widehat{e})^2] \tag{11}$$

to estimate the variance, where $\widehat{\epsilon}_i$ are the i-th residuals.

Theorem 10 Let $(\widetilde{x}_{i1},\widetilde{x}_{i2},\cdots,\widetilde{x}_{ip},\widehat{z}_i,\widetilde{y}_i)$, $i=1,2,\cdots,n$, be a set of imprecisely observed data, where $\widetilde{x}_{i1},\widetilde{x}_{i2},\cdots,\widetilde{x}_{ip},\widetilde{z}_i,\widetilde{y}_i$ are independent uncertain variables with regular uncertainty distributions $\Gamma_{i1}, \Gamma_{i2}, \cdots, \Gamma_{ip}, \Phi_i, \Psi_i$, respectively, and let the fitted semiparametric regression model be

$$y = \sum_{j=1}^{p} \widehat{\beta}_j x_j + \widehat{f}(z). \tag{12}$$

Then the estimated expected value of the disturbance term ϵ is

$$\widehat{e} = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \Big[\Psi_i^{-1}(\alpha) - \sum_{j=1}^{p} \widehat{\beta}_j \Pi_{ij}^{-1}(\alpha, \widehat{\beta}_j) - \widehat{f}(\Upsilon_i^{-1}(\alpha, \widehat{f}))\Big] \mathrm{d}\alpha, \tag{13}$$

and the estimated variance is

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \Big[\Psi_i^{-1}(\alpha) - \sum_{j=1}^{p} \widehat{\beta}_j \Pi_{ij}^{-1}(\alpha, \widehat{\beta}_j) - \widehat{f}(\Upsilon_i^{-1}(\alpha, \widehat{f})) - \widehat{e}\Big]^2 \mathrm{d}\alpha, \tag{14}$$

where

$$\Pi_{ij}^{-1}(\alpha, \widehat{\beta}_j) = \begin{cases} \Gamma_{ij}^{-1}(1-\alpha), & \text{if } \widehat{\beta}_j > 0; \\ \Gamma_{ij}^{-1}(\alpha), & \text{if } \widehat{\beta}_j < 0, \end{cases}$$

and

$$\Upsilon_i^{-1}(\alpha, \widehat{f}) = \begin{cases} \Phi_i^{-1}(1-\alpha), & \text{if } \widehat{f} \text{ is strictly increasing;} \\ \Phi_i^{-1}(\alpha), & \text{if } \widehat{f} \text{ is strictly decreasing,} \end{cases}$$

for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$.

Proof For each i, it follows from [22] that the inverse uncertainty distribution of $\widetilde{y}_i - \sum\limits_{j=1}^{p} \widehat{\beta}_j \widetilde{x}_j - \widehat{f}(\widetilde{z}_i)$ is

$$F_i^{-1}(\alpha) = \Psi_i^{-1}(\alpha) - \sum_{j=1}^{p} \widehat{\beta}_j \Pi_{ij}^{-1}(\alpha, \widehat{\beta}_j) - \widehat{f}(\Upsilon_i^{-1}(\alpha, \widehat{f})).$$

Similarly to the proof of Theorem 8, the theorem follows immediately. □

Suppose $(\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_p, \widetilde{z})$ is a vector of new explanatory variables, where $\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_p, \widetilde{z}$ are uncertain variables with regular uncertainty distributions $\Gamma_1, \Gamma_2, \cdots, \Gamma_p, \Phi$, respectively. Based on the fitted model, we can forecast the response variable for a new explanatory vector. Now suppose the fitted semiparametric regression model is

$$y = \sum_{j=1}^{p} \widehat{\beta}_j x_j + \widehat{f}(z),$$

the disturbance term ϵ is independent of $\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_p, \widetilde{z}$, and its estimated expected value and variance are $\widehat{e}$ and $\widehat{\sigma}^2$, respectively. Then the forecast uncertain variable of y with respect to $(\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_p, \widetilde{z})$ is

$$\widehat{y} = \sum_{j=1}^{p} \widehat{\beta}_j \widetilde{x}_j + \widehat{f}(\widetilde{z}) + \epsilon.$$

Note that y is an uncertain variable, and we define the expected value of the forecast uncertain variable $\widehat{y}$ as the forecast value of y,

$$\mu = \sum_{j=1}^{p} \widehat{\beta}_j \mathsf{E}[\widetilde{x}_j] + \mathsf{E}[f(\widetilde{z})] + \widehat{e}. \tag{15}$$

§4. Numerical Example

Assume the social benefits of a new factory is studied based on experts' data and decide whether to establish the new factory. There are some factors that affect social benefits such as quality of the production and carbon emission. We take social benefits (y) as a response variable, quality of the production (x) and carbon emission (z) as explanatory variables. Due to the data cannot be precisely observed, we treat x, z, y as uncertain variables. The relationship between y and x is linear by general knowledge. The relationship between y and z is complicated and taken as nonlinear effect. In our analysis, we believe that it is reasonable to assume that the social benefits are decreasing with respect to the carbon emission, since carbon emission has devastating effect on social benefits because of environmental pollution. Thus, we employ the uncertain semiparametric regression model

$$\widetilde{y}_i = \widetilde{x}_i \beta + f(\widetilde{z}_i) + \epsilon_i, \qquad i = 1, 2, \cdots, 26,$$

where f is monotone decreasing function. Suppose the data $(\widetilde{x}_i, \widetilde{z}_i, \widetilde{y}_i)$, $i = 1, 2, \cdots, 26$, are acquired by using the questionnaire survey, see Table 1, where $\mathscr{L}(a, b)$ denotes linear uncertain variable. For each i, $\widetilde{x}_i$, $\widehat{z}_i$ and $\widetilde{y}_i$ are independent.

We use monotone Bernstein polynomials to approximate the nonparametric function f. The order m is selected by V-fold cross-validation method. It is one of the most widely used methods to estimate prediction error. Given V, for each m, the cross-validation term $\mathrm{CV}(m)$ takes the following form:

$$\mathrm{CV}(m) = \frac{1}{V} \sum_{i=1}^{V} \sum_{i \in I_{-v}} \mathsf{E}\big[\widetilde{y}_i - \widetilde{x}_i \widehat{\beta}_{m(v)} - \widehat{f}_{m(v)}(\widetilde{z}_i)\big]^2,$$

where $\widehat{\beta}_{m(v)}$ and $\widehat{f}_{m(v)}$ obtained from the v-th training data set consisting of $\lfloor n(V-1)/V \rfloor$ observation points and I_{-v} denotes the corresponding validation set consisting of $[n/V]$ points. We compute the cross-validation function $\mathrm{CV}(m)$ for a series of m values starting with $m = 2$ to a relatively large integer $(\lfloor n(V-1)/V-1 \rfloor)$. The optimal value $\widehat{m}$ is chosen to minimize $\mathrm{CV}(m)$, i.e., $\widehat{m} = \underset{m}{\operatorname{argmin}}\, \mathrm{CV}(m)$. we take $V = 1$ and optimal value is $\widehat{m} = 6$. For more details on cross-validation method in uncertain statistics, the reader can refer to

Table 1 Imprecisely observed data

i	$\widetilde{x}_i$	$\widetilde{z}_i$	$\widetilde{y}_i$
1	$\mathscr{L}(5, 6.8)$	$\mathscr{L}(1, 2.4)$	$\mathscr{L}(1100, 1144)$
2	$\mathscr{L}(5.2, 6)$	$\mathscr{L}(3, 4.2)$	$\mathscr{L}(700, 802)$
3	$\mathscr{L}(8, 9.4)$	$\mathscr{L}(1, 3)$	$\mathscr{L}(900, 1118)$
4	$\mathscr{L}(6, 6.8)$	$\mathscr{L}(3, 3.2)$	$\mathscr{L}(800, 1034)$
5	$\mathscr{L}(6, 6.6)$	$\mathscr{L}(3, 4.6)$	$\mathscr{L}(900, 1072)$
6	$\mathscr{L}(9, 11.2)$	$\mathscr{L}(3, 3.8)$	$\mathscr{L}(800, 938)$
7	$\mathscr{L}(8, 11.2)$	$\mathscr{L}(2, 2.6)$	$\mathscr{L}(806, 1006)$
8	$\mathscr{L}(7, 8.4)$	$\mathscr{L}(3, 4.8)$	$\mathscr{L}(800, 1062)$
9	$\mathscr{L}(8, 9.6)$	$\mathscr{L}(2, 3.8)$	$\mathscr{L}(800, 916)$
10	$\mathscr{L}(8, 8.6)$	$\mathscr{L}(2, 2.4)$	$\mathscr{L}(1000, 1254)$
11	$\mathscr{L}(6, 9.4)$	$\mathscr{L}(1.9, 2.5)$	$\mathscr{L}(1000, 1448)$
12	$\mathscr{L}(7, 8.2)$	$\mathscr{L}(2.9, 3.3)$	$\mathscr{L}(700, 754)$
13	$\mathscr{L}(6, 7.2)$	$\mathscr{L}(3, 4.2)$	$\mathscr{L}(800, 926)$
14	$\mathscr{L}(8, 9.6)$	$\mathscr{L}(4, 4.6)$	$\mathscr{L}(700, 840)$
15	$\mathscr{L}(8, 9.4)$	$\mathscr{L}(2, 3.8)$	$\mathscr{L}(900, 1040)$
16	$\mathscr{L}(7, 8.6)$	$\mathscr{L}(2.5, 3.3)$	$\mathscr{L}(900, 944)$
17	$\mathscr{L}(9, 10)$	$\mathscr{L}(3, 5)$	$\mathscr{L}(700, 976)$
18	$\mathscr{L}(7, 8)$	$\mathscr{L}(4, 5.2)$	$\mathscr{L}(900, 964)$
19	$\mathscr{L}(6, 7.6)$	$\mathscr{L}(2, 3.4)$	$\mathscr{L}(800, 876)$
20	$\mathscr{L}(7, 7.6)$	$\mathscr{L}(3, 3.4)$	$\mathscr{L}(750, 1048)$
21	$\mathscr{L}(8, 8.6)$	$\mathscr{L}(2, 3.8)$	$\mathscr{L}(700, 900)$
22	$\mathscr{L}(8, 8.8)$	$\mathscr{L}(3, 4.8)$	$\mathscr{L}(900, 1024)$
23	$\mathscr{L}(7, 8.2)$	$\mathscr{L}(2, 2.8)$	$\mathscr{L}(800, 1172)$
24	$\mathscr{L}(8, 9.2)$	$\mathscr{L}(2, 2.4)$	$\mathscr{L}(900, 1138)$
25	$\mathscr{L}(7, 7.2)$	$\mathscr{L}(1, 3)$	$\mathscr{L}(850, 1312)$
26	$\mathscr{L}(9, 11)$	$\mathscr{L}(2, 3.4)$	$\mathscr{L}(700, 890)$

[25] and [26]. By solving the minimization problem (9), we get the least squares estimate of β is

$$\widehat{\beta} = 101.7,$$

and the least squares estimate of $\boldsymbol{b}$ is

$$\widehat{\boldsymbol{b}} = (1529.7, -34.8, 12.1, 59, 105.9, 153.6, 317.7).$$

The fitted nonparametric function is $\widehat{f}(z) = \boldsymbol{\delta}(z)\widehat{\boldsymbol{b}}$, and the fitted curve of $\widehat{f}(z)$ is shown in Figure 1, which shows the social benefits are decreasingly with respect to the carbon emission. By using the formulas (13) and (14), we get the expected value and variance of the disturbance term ϵ are

$$\widehat{e} = -0.014, \qquad \widehat{\sigma}^2 = 19804,$$

respectively. Now let $(\widetilde{x}, \widetilde{z}) = (\mathscr{L}(10, 12), \mathscr{L}(4, 6))$ is a new vector of explanatory variables, by calculating the formula (15), the forecast value of response variable y is $\mu = 1175$.

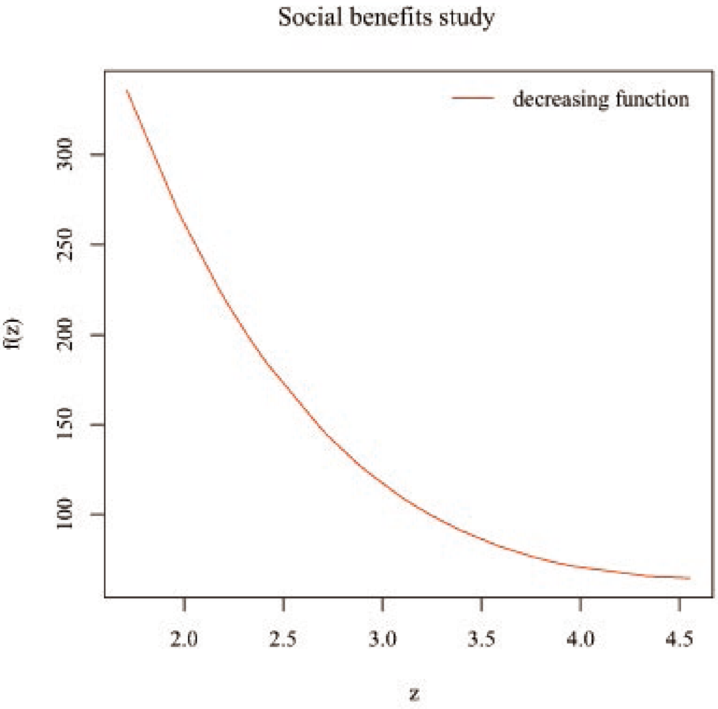

Figure 1 The estimate of function f

§5. Conclusions

This paper proposes the least squares estimate of the uncertain semiparametric regression model when the nonparametric component is subject to monotonicity constraint based on the uncertainty theory. By employing Bernstain polynomials, semiparametric regression model are turned into the linear regression model framework. Quadratic programming algorithm is employed to compute the estimate. The cross validation method is used to select the number of basis functions. A numerical example shows the proposed methods are effective.

References

[1] YAO K. Uncertain statistical inference models with imprecise observations [J]. *IEEE T Fuzzy Syst*, 2018, **26(2)**: 409–415.

[2] WEN M L, ZHANG Q Y, KANG R, et al. Some new ranking criteria in data envelopment analysis under uncertain environment [J]. *Comput Ind Eng*, 2017, **110**: 498–504.

[3] WANG X S, GAO Z C, GUO H Y. Delphi method for estimating uncertainty distributions [J]. *Information*, 2012, **15(2)**: 449–459.

[4] ZHAO X, PENG J, LIU J, et al. Analytic solution of uncertain autoregressive model based on principle of least squares [J]. *Soft Comput*, 2020, **24(4)**: 2721–2726.

[5] LIO W, LIU B D. Uncertain data envelopment analysis with imprecisely observed inputs and outputs [J]. *Fuzzy Optim Decis Mak*, 2018, **17(3)**: 357–373.

[6] LIU B D. Uncertain risk analysis and uncertain reliability analysis [J]. *J Uncertain Syst*, 2010, **4(3)**: 163–170.

[7] LIU B D. *Theory and Practice of Uncertain Programming* [M]. 2nd ed. Berlin: Springer, 2009.

[8] YAO K, LI X. Uncertain alternating renewal process and its application [J]. *IEEE T Fuzzy Syst*, 2012, **20(6)**: 1154–1160.

[9] LIU B D. *Uncertainty Theory: A Branch of Mathematics for Modeling Human Uncertainty* [M]. Berlin: Springer, 2010.

[10] YAO K, LIU B D. Uncertain regression analysis: an approach for imprecise observations [J]. *Soft Comput*, 2018, **22(17)**: 5579–5582.

[11] LIO W, LIU B D. Residual and confidence interval for uncertain regression model with imprecise observations [J]. *J Intell Fuzzy Syst*, 2018, **35(2)**: 2573–2583.

[12] LIO W, LIU B D. Uncertain maximum likelihood estimation with application to uncertain regression analysis [J]. *Soft Comput*, 2020, **24(13)**: 9351–9360.

[13] SONG Y L, FU Z F. Uncertain multivariable regression model [J]. *Soft Comput*, 2018, **22(17)**: 5861–5866.

[14] FANG L, HONG Y P. Uncertain revised regression analysis with responses of logarithmic, square root and reciprocal transformations [J]. *Soft Comput*, 2020, **24(4)**: 2655–2670.

[15] YANG X F, NI Y D. Least-squares estimation for uncertain moving average model [J]. *Comm Statist Theory Methods*, 2021, **50(17)**: 4134–4143.

[16] LIU Z, YANG Y. Least absolute deviations estimation for uncertain regression with imprecise observations [J]. *Fuzzy Optim Decis Mak*, 2020, **19(1)**: 33–52.

[17] LIU Z, JIA L F. Cross-validation for the uncertain Chapman-Richards growth model with imprecise observations [J]. *Internat J Uncertain Fuzziness Knowledge-Based Systems*, 2020, **28(5)**: 769–783.

[18] DING J H, ZHANG Z Q. Statistical inference on uncertain nonparametric regression model [J]. *Fuzzy Optim Decis Mak*, 2021, **20(4)**: 451–469.

[19] LIU B D. *Uncertainty Theory* [M]. 2nd ed. Berlin: Springer, 2007.

[20] LIU B D. Some research problems in uncertainty theory [J]. *J Uncertain Syst*, 2009, **3(1)**: 3–10.

[21] LORENTZ G G. *Bernstein Polynomials* [M]. 2nd ed. New York: Chelsea Publishing Company, 1986.

[22] LIU B D. *Uncertainty Theory* [M]. 4th ed. Heidelberg: Springer, 2015.

[23] GOLDFARB D, IDNANI A. A numerically stable dual method for solving strictly convex quadratic programs [J]. *Math Programming*, 1983, **27(1)**: 1–33.

[24] TURLACH B A, WEINGESSEL A. Quadprog (Version 1.5-3), R package [CP/DK]. 2010.

[25] FANG L, LIU S Q, HUANG Z Y. Uncertain Johnson-Schumacher growth model with imprecise observations and k-fold cross-validation test [J]. *Soft Comput*, 2020, **24(4)**: 2715–2720.

[26] LIU Z, YANG X F. Cross validation for uncertain autoregressive model [J/OL]. *Comm Statist Simulation Comput*, 2020 [2020-4-6]. https://doi.org/10.1080/03610918.2020.1747077.

不确定半参数模型的 Bernstein 多项式估计

丁建华[1,2]　张红玉[1]　张志强[1]

([1]山西大同大学统计系, 大同, 037009)

([2]统计与数据科学前沿理论及应用教育部重点实验室, 上海, 200062)

摘　要: 假设样本观测值是不精确的, 通过将不精确的观测值建模为不确定变量, 这篇论文提出单调半参数模型的不确定统计推断. 单调 Bernstein 多项式近似非参数函数, 利用二次规划算法进行求解. 并通过数值例子说明所提出的方法.

关键词: 不确定变量; 不确定半参数模型; Bernstein 多项式; 交叉证实法

中图分类号: O212.7