Soft Comput
DOI 10.1007/s00500-017-2521-y

FOCUS

Uncertain regression analysis: an approach for imprecise observations

Kai Yao[1,2] · Baoding Liu[3]

Abstract Regression analysis is a method to estimate the relationships among the response variable and the explanatory variables. Assuming the observations of the response variable are imprecise and modeling the observed data via uncertain variables, this paper explores an approach of uncertain regression analysis to estimating the relationships among the variables with imprecisely observed samples. On the principle of least squares, an optimization problem is derived to calculate the unknown parameters in the regression model. In particular, this paper investigates uncertain linear regression model and gives an analytic representation of the unknown parameters.

Keywords Uncertain regression · Uncertain linear regression · Uncertain variable · Uncertainty theory

1 Introduction

Probability theory has been used to deal with the indeterminate quantities for a long time. A premise of applying probability theory is that the obtained probability distribution is close enough to the frequency of the indeterminate quantity. But in practice, we are often in the predicament of deficient in observed data. In this case, we have to invite the domain experts to evaluate the possibilities that some events occur. Since *human beings usually overweight unlikely events* (Kahneman and Tversky 1979) and *human beings usually estimate a much wider range of values than the object actually takes* (Liu 2015), the belief degree generally cannot be regarded as a probability distribution. Thus a new approach to dealing with belief degree is needed.

Uncertainty theory was founded by Liu (2007) and perfected by Liu (2009) based on normality, duality, subadditivity, and product axioms. It aims at dealing with the belief degree and the quantity with human uncertainty. In the framework of uncertainty theory, an uncertain measure is used to indicate the belief degree, and an uncertain variable is used to model the quantity with human uncertainty. Besides, a concept of uncertainty distribution is employed to describe an uncertain variable, and concepts of expected value, variance, and entropy are employed to characterize an uncertain variable.

In practice, the uncertainty distribution of an uncertain variable usually comes from the belief degree of the experts. To collect the expert's experimental data and produce the empirical uncertainty distribution, Liu (2010) designed a questionnaire survey in the form "how likely is ξ not more than x." Chen and Ralescu (2012) applied this questionnaire survey to estimating the distance from Tianjin to Beijing, and the results showed the questionnaire survey is highly effective. In the case of multiple experts, Wang et al. (2012) recast the Delphi method to determine the uncertainty distributions. So far, uncertain statistics has branches into parametric uncertain statistics and nonparametric uncertain statistics. For the former one, Liu (2010) suggested the principle of least

Communicated by Y. Ni.

✉ Kai Yao
yaokai@ucas.ac.cn

Baoding Liu
liu@tsinghua.edu.cn

[1] School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China

[2] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China

[3] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

Published online: 15 February 2017

squares to estimate the unknown parameters in the uncertainty distribution, and Wang and Peng (2014) suggested the method of moments. For the latter one, Liu (2010) employed the linear interpolation method to generate the uncertainty distribution from expert's experimental data, and Chen and Ralescu (2012) employed the B-spline interpolation method.

Note that the traditional regression analysis estimates the relationships among the response variable and the explanatory variables based on the assumption that the samples of these variables are precisely observed. But due to the perturbation, the observations of the samples are sometimes imprecise. In this paper, we model the samples of the response variable via uncertain variables and found an uncertain regression model to estimate the relationships among the variables with imprecisely observed samples. The rest of this paper is organized as follows. In Sect. 2, we introduce some basic results about uncertain variables. Then a stipulation about the square of an uncertain variable is made in Sect. 3. After that, uncertain regression analysis and uncertain linear regression analysis are proposed in Sects. 4 and 5, respectively. Finally, some remarks are made in Sect. 6.

2 Preliminaries

Uncertain measure is a function defined on a σ-algebra, and it is used to indicate human's belief degree.

Definition 1 (Liu 2007) Let $\mathcal{L}$ be a σ-algebra on a nonempty set Γ. A set function $\mathcal{M} : \mathcal{L} \to [0, 1]$ is called an uncertain measure if it satisfies the following axioms:

Axiom 1: (Normality Axiom) $\mathcal{M}\{\Gamma\} = 1$ *for the universal set* Γ.
Axiom 2: (Duality Axiom) $\mathcal{M}\{\Lambda\} + \mathcal{M}\{\Lambda^c\} = 1$ *for any event* Λ.
Axiom 3: (Subadditivity Axiom) For every countable sequence of events $\Lambda_1, \Lambda_2, \ldots,$ *we have*

$$\mathcal{M}\left\{\bigcup_{i=1}^{\infty} \Lambda_i\right\} \le \sum_{i=1}^{\infty} \mathcal{M}\{\Lambda_i\}.$$

In this case, the triple $(\Gamma, \mathcal{L}, \mathcal{M})$ is called an uncertainty space.

The product uncertain measure on the product σ-algebra of the uncertain spaces was defined by Liu (2009), producing the fourth axiom of uncertain measure.
Axiom 4: (Product Axiom) Let $(\Gamma_k, \mathcal{L}_k, \mathcal{M}_k)$ *be uncertainty spaces for* $k = 1, 2, \ldots$ *Then the product uncertain measure* $\mathcal{M}$ *is an uncertain measure satisfying*

$$\mathcal{M}\left\{\prod_{k=1}^{\infty} \Lambda_k\right\} = \bigwedge_{k=1}^{\infty} \mathcal{M}_k\{\Lambda_k\}$$

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

As a real function on the uncertainty space, uncertain variable is used to model the quantities with human uncertainty.

Definition 2 (Liu 2007) An uncertain variable ξ is a measurable function from the uncertainty space $(\Gamma, \mathcal{L}, \mathcal{M})$ to the set of real numbers, i.e., for any Borel set B of real numbers, the set

$$\{\xi \in B\} = \{\gamma \in \Gamma | \xi(\gamma) \in B\}$$

is an event.

Definition 3 (Liu 2009) The uncertain variables $\xi_1, \xi_2, \ldots, \xi_n$ are said to be independent if

$$\mathcal{M}\left\{\bigcap_{i=1}^{n} (\xi_i \in B_i)\right\} = \bigwedge_{i=1}^{n} \mathcal{M}\{\xi_i \in B_i\} \tag{1}$$

for any Borel sets $B_1, B_2, \ldots, B_n$ of real numbers.

Uncertainty distribution is used to describe an uncertain variable, but it carries only partial information about an uncertain variable.

Definition 4 (Liu 2007) The uncertainty distribution of an uncertain variable ξ is defined by

$$\Phi(x) = \mathcal{M}\{\xi \le x\}$$

for any real number x.

Assume that $\xi_1, \xi_2, \ldots, \xi_n$ are independent uncertain variables with continuous uncertainty distributions $\Phi_1, \Phi_2, \ldots, \Phi_n$, respectively. Liu (2009) showed that if the function $f(x_1, x_2, \ldots, x_n)$ is strictly increasing with respect to $x_1, x_2, \ldots, x_m$ and strictly decreasing with respect to $x_{m+1}, x_{m+2}, \ldots, x_n$, then the uncertain variable

$$\xi = f(\xi_1, \xi_2, \ldots, \xi_n)$$

has an uncertainty distribution

$$\Phi(x) = \sup_{f(x_1, x_2, \ldots, x_n) \le x} \left(\min_{1 \le i \le m} \Phi_i(x_i) \wedge \min_{m+1 \le i \le n} (1 - \Phi_i(x_i)) \right).$$

Definition 5 (Liu 2007) The expected value of an uncertain variable ξ is defined by

$$E[\xi]=\int_0^{+\infty}\mathcal{M}\{\xi\ge x\}\mathrm{d}x-\int_{-\infty}^0\mathcal{M}\{\xi\le x\}\mathrm{d}x$$

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 $E[\xi]$ exists, then

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

3 Stipulation about uncertain variables

Note that an uncertainty distribution carries only partial information about an uncertain variable, so some characteristics of the uncertain variable cannot be precisely derived from its uncertainty distribution. In this case, we accept a stipulation about uncertain variables as below.

Consider an uncertain variable ξ with an uncertainty distribution Φ. Given a real number r, we are interested in $E\left[(\xi-r)^2\right]$ which is frequently used in uncertain statistics. By using the duality and subadditivity of uncertain measure, we have

$$\begin{aligned}E\left[(\xi-r)^2\right]&=\int_0^{+\infty}\mathcal{M}\left\{(\xi-r)^2\ge x\right\}\mathrm{d}x\\&=\int_0^{+\infty}\mathcal{M}\left\{\left(\xi\ge r+\sqrt{x}\right)\cup\left(\xi\le r-\sqrt{x}\right)\right\}\mathrm{d}x\\&\le\int_0^{+\infty}\left(\mathcal{M}\left\{\xi\ge r+\sqrt{x}\right\}+\mathcal{M}\left\{\xi\le r-\sqrt{x}\right\}\right)\mathrm{d}x\\&=\int_0^{+\infty}\left(1-\Phi\left(r+\sqrt{x}\right)\right)\mathrm{d}x+\int_0^{+\infty}\Phi\left(r-\sqrt{x}\right)\mathrm{d}x.\end{aligned}$$

In the first term, substituting $r+\sqrt{x}$ with y and x with $(y-r)^2$, we have

$$\begin{aligned}\int_0^{+\infty}\left(1-\Phi\left(r+\sqrt{x}\right)\right)\mathrm{d}x&=\int_r^{+\infty}(1-\Phi(y))\,\mathrm{d}(y-r)^2\\&=\int_r^{+\infty}(y-r)^2\mathrm{d}\Phi(y).\end{aligned}$$

In the second term, substituting $r-\sqrt{x}$ with y and x with $(r-y)^2$, we have

$$\begin{aligned}\int_0^{+\infty}\Phi\left(r-\sqrt{x}\right)\mathrm{d}x&=\int_r^{-\infty}\Phi(y)\mathrm{d}(r-y)^2\\&=\int_{-\infty}^r(y-r)^2\mathrm{d}\Phi(y).\end{aligned}$$

As a result,

$$\begin{aligned}E\left[(\xi-r)^2\right]&\le\int_r^{+\infty}(y-r)^2\mathrm{d}\Phi(y)+\int_{-\infty}^r(y-r)^2\mathrm{d}\Phi(y)\\&=\int_{-\infty}^{+\infty}(y-r)^2\mathrm{d}\Phi(y).\end{aligned}$$

In this case, we stipulate

$$E\left[(\xi-r)^2\right]=\int_{-\infty}^{+\infty}(y-r)^2\mathrm{d}\Phi(y).\tag{2}$$

In the stipulation (2), taking $r=E[\xi]$, we have

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

Furthermore,

$$\begin{aligned}E\left[(\xi-r)^2\right]&=\int_{-\infty}^{+\infty}(y-E[\xi]+E[\xi]-r)^2\mathrm{d}\Phi(y)\\&=\int_{-\infty}^{+\infty}(y-E[\xi])^2\mathrm{d}\Phi(y)+(E[\xi]-r)^2\\&\quad+2(E[\xi]-r)\cdot\int_{-\infty}^{+\infty}(y-E[\xi])\mathrm{d}\Phi(y)\\&=V[\xi]+(E[\xi]-r)^2.\end{aligned}$$

4 Uncertain regression analysis

Let $\boldsymbol{x}=(x_1,x_2,\ldots,x_p)$ be a vector of explanatory variables, and let y be a response variable. Assume the functional relationship between y and $\boldsymbol{x}$ can be expressed by the regression model

$$y=f(\boldsymbol{x}|\boldsymbol{\beta})+\varepsilon\tag{3}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters, and ε is a disturbance term.

Traditionally, it is assumed that both $\boldsymbol{x}$ and y are able to be precisely observed. However, in many cases, the observations of those variables are imprecise and characterized in terms of uncertain variables. Now suppose that we have a set of observed data,

$$(\boldsymbol{x}_1,\tilde{y}_1),\ (\boldsymbol{x}_2,\tilde{y}_2),\ \ldots,(\boldsymbol{x}_n,\tilde{y}_n)\tag{4}$$

where $\boldsymbol{x}_i$ are crisp explanatory vectors, and $\tilde{y}_i$ are uncertain response variables for $i=1,2,\ldots,n$.

Based on the observed data $(\boldsymbol{x}_1,\tilde{y}_1),(\boldsymbol{x}_2,\tilde{y}_2),\ldots,(\boldsymbol{x}_n,\tilde{y}_n)$, the least squares estimate of $\boldsymbol{\beta}$ in the regression model

(3) is the solution of the minimization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} E[(\tilde{y}_i - f(\boldsymbol{x}_i|\boldsymbol{\beta}))^2]. \tag{5}$$

Denote the optimal solution by $\boldsymbol{\beta}^*$, then the fitted regression model is determined by $y = f(\boldsymbol{x}|\boldsymbol{\beta}^*)$.

Theorem 1 *Let* $(\boldsymbol{x}_1, \tilde{y}_1)$, $(\boldsymbol{x}$