Fuzzy Optimization and Decision Making (2022) 21:157–174
https://doi.org/10.1007/s10700-021-09365-w

Uncertain hypothesis test with application to uncertain regression analysis

Tingqing Ye[1] · Baoding Liu[1]

Accepted: 19 June 2021 / Published online: 9 July 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
This paper first establishes uncertain hypothesis test as a mathematical tool that uses uncertainty theory to help people rationally judge whether some hypotheses are correct or not, according to observed data. As an application, uncertain hypothesis test is employed in uncertain regression analysis to test whether the estimated disturbance term and the fitted regression model are appropriate. In order to illustrate the test process, some numerical examples are documented.

Keywords Uncertainty theory · Uncertain statistics · Hypothesis test · Regression analysis

1 Introduction

Uncertainty theory, founded by Liu (2007) and perfected by Liu (2009), has been a branch of mathematics and successfully applied in the fields like science and engineering. As an important application of uncertainty theory, uncertain statistics, first discussed by Liu (2010), is a methodology of collecting, analyzing and interpreting data based on uncertainty theory. Up to now, uncertain statistics has four main development fields: estimating uncertainty distribution, uncertain regression analysis, uncertain time series analysis, and parameter estimation in uncertain differential equation.

Estimating uncertainty distribution is aimed to use uncertainty theory to fit the uncertainty distribution for an uncertain variable based on the expert's experimental data. The first step to estimate uncertainty distribution is to collect expert's experimental data. For that matter, Liu (2010) designed a questionnaire survey. The next step

✉ Baoding Liu
liu@tsinghua.edu.cn

Tingqing Ye
yetq18@mails.tsinghua.edu.cn

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

is to fit the uncertainty distribution based on the collected expert's experimental data. If the functional form of the uncertainty distribution to be estimated is known but it contains some unknown parameters, then in order to estimate the unknown parameters via expert's experimental data, Liu (2010) investigated the principle of least squares, and Wang and Peng (2014) presented the method of moments. If the functional form of the uncertainty distribution to be estimated is even unknown, then in order to estimate the uncertainty distributions, Liu (2010) presented the linear interpolation method, and Chen and Ralescu (2012) explored a series of spline interpolation methods. In addition, Delphi method (Wang et al. (2012)) was suggested as a process to estimate the uncertainty distribution when multiple experts are available.

Uncertain regression analysis is aimed to use uncertainty theory to study the relationship between explanatory variables and response variables. Parameter estimation for the unknown parameters is a vital topic in uncertain regression analysis. Many approaches about estimating unknown parameters in uncertain regression models have been developed, such as the least squares estimation (Yao and Liu (2018)), the least absolute deviations estimation (Liu and Yang (2020)), and the maximum likelihood estimation (Lio and Liu (2020)). In addition, Lio and Liu (2018) proposed an approach to make interval estimation for predicting the response variables. Furthermore, there are many other directions in uncertain regression analysis, including cross-validation (Liu and Jia (2020); Liu (2019)), variable selection (Liu and Yang (2020)), multivariate regression analysis (Song and Fu (2018); Ye and Liu (2020)), and nonparametric regression analysis (Ding and Zhang (2021)).

Uncertain time series analysis is aimed to use uncertainty theory to predict future values based on previously observed data. As a basic model of uncertain time series, uncertain autoregressive model was first proposed by Yang and Liu (2019). In the uncertain autoregressive model, the observed data depend linearly on its previous values and an uncertain disturbance term. In order to take the multiple uncertain disturbance terms in uncertain time series into consideration, Yang and Ni (2020) presented uncertain moving average model where the observed data depend linearly on the current and various past values of a disturbance term.

Parameter estimation in uncertain differential equation is aimed to use uncertainty theory to estimate unknown parameters in uncertain differential equation based on observed data. Many researchers have studied lots of methods of parameter estimation in uncertain differential equation. For example, Yao and Liu (2020) investigated moment estimation, Yang et al. (2020) studied minimum cover estimation, Sheng et al. (2021) investigated least squares estimation, Liu (2021) proposed generalized moment estimation, and Liu and Liu (2020) presented maximum likelihood estimation. As another topic, initial value estimation was proposed by Lio and Liu (2021) to estimate the unknown initial value of uncertain differential equation according to observed data.

This paper explores to develop a new direction of uncertain statistics called uncertain hypothesis test, which is concerned with using uncertainty theory to make decisions about whether some hypotheses are correct or not, according to observed data. As a purpose of investigating uncertain hypothesis test, we employ it in uncertain regression analysis to test whether the estimated disturbance term and the fitted regression model are appropriate.

The rest of the paper is organized as follows. Uncertain hypothesis test is introduced in Sect. 2, and is applied in uncertain regression analysis in Sect. 3. Then, some numerical examples are given in Sect. 4. Finally, a brief summary is made in Sect. 5.

2 Uncertain hypothesis test

Let ξ be a population with uncertainty distribution Φ_θ where θ is an unknown parameter with $\theta \in \Theta$. A hypothesis testing problem about the unknown parameter θ can be formulated as deciding which of the following two statements is true:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1 \tag{1}$$

where Θ_0 and Θ_1 are two disjoint subsets of Θ and $\Theta_0 \cup \Theta_1 = \Theta$. The statement H_0 is called a null hypothesis, and H_1 is called an alternative hypothesis. Especially, the following hypotheses are called two-sided hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0,$$

where $\theta_0 \in \Theta$.

Assume there is a vector of observed data $(z_1, z_2, \cdots, z_n)$. A rejection region for the null hypothesis H_0 is a set $W \subset \Re^n$. If the vector of observed data

$$(z_1, z_2, \cdots, z_n) \in W,$$

then we reject H_0. Otherwise, we accept H_0. A core problem is how to choose a suitable rejection region W for the given hypothesis H_0.

Definition 1 Let ξ be a population with uncertainty distribution Φ_θ where θ is an unknown parameter. A rejection region $W \subset \Re^n$ is said to be a test for the two-sided hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ at significance level α if

(a) for any $(z_1, z_2, \cdots, z_n) \in W$, there are at least α of indexes i's with $1 \le i \le n$ such that

$$\mathcal{M}_{\theta_0}\{\xi > z_i\} \vee \mathcal{M}_{\theta_0}\{\xi < z_i\} > 1 - \frac{\alpha}{2},$$

(b) for some $\theta \neq \theta_0$ and some $(z_1, z_2, \cdots, z_n) \in W$, there are more than $1 - \alpha$ of indexes i's with $1 \le i \le n$ and at least α of indexes j's with $1 \le j \le n$ such that

$$\mathcal{M}_{\theta}\{\xi > z_i\} \vee \mathcal{M}_{\theta}\{\xi < z_i\} < \mathcal{M}_{\theta_0}\{\xi > z_j\} \vee \mathcal{M}_{\theta_0}\{\xi < z_j\}.$$

Remark 1 From Definition 1, we can see that the test W is related to the significance level α. How do we choose it? Standard values, such as 0.1, 0.05, or 0.01, are often used for convenience.

In order to find a suitable rejection region W satisfying the two conditions in Definition 1, we introduce a concept of nonembedded uncertainty distribution family.

Definition 2 A regular uncertainty distribution family $\{\Phi_\theta : \theta \in \Theta\}$ is said to be nonembedded for $\theta_0 \in \Theta$ at level α if

$$\Phi_{\theta_0}^{-1}(\beta) > \Phi_\theta^{-1}(\beta) \quad \text{or} \quad \Phi_\theta^{-1}(1-\beta) > \Phi_{\theta_0}^{-1}(1-\beta)$$

for some $\theta \in \Theta$ and some β with $0 < \beta \le \alpha/2$.

Example 1 The normal uncertainty distribution family $\{\mathcal{N}(e,\sigma) : e \in \Re, \sigma > 0\}$ is nonembedded for any $\theta_0 = (e_0, \sigma_0) \in \Re \times (0, +\infty)$ at any level α. Note that the inverse uncertainty distribution of $\mathcal{N}(e,\sigma)$ is

$$\Phi^{-1}(\beta) = e + \frac{\sigma\sqrt{3}}{\pi} \ln \frac{\beta}{1-\beta}.$$

Take

$$\theta_1 = (e_1, \sigma_1) = (e_0 - 1, \sigma_0), \quad \beta = \frac{\alpha}{2}.$$

Since

$$\begin{aligned}\Phi_{\theta_0}^{-1}(\beta) - \Phi_{\theta_1}^{-1}(\beta) &= e_0 + \frac{\sigma_0\sqrt{3}}{\pi} \ln \frac{\beta}{1-\beta} - \left(e_1 + \frac{\sigma_1\sqrt{3}}{\pi} \ln \frac{\beta}{1-\beta}\right) \\ &= e_0 + \frac{\sigma_0\sqrt{3}}{\pi} \ln \frac{\beta}{1-\beta} - \left(e_0 - 1 + \frac{\sigma_0\sqrt{3}}{\pi} \ln \frac{\beta}{1-\beta}\right) \\ &= 1 > 0,\end{aligned}$$

the normal uncertainty distribution family $\{\mathcal{N}(\theta,\sigma) : e \in \Re, \sigma > 0\}$ is nonembedded for θ_0 at level α.

Example 2 The linear uncertainty distribution family $\{\mathcal{L}(a,b) : a < b\}$ is nonembedded for any $\theta_0 = (a_0, b_0)$ with $a_0 < b_0$ at any level α. Note that the inverse uncertainty distribution of $\mathcal{L}(a,b)$ is

$$\Phi^{-1}(\beta) = (1-\beta)a + \beta\theta.$$

Take

$$\theta_1 = (a_1, b_1) = (a_0 - 1, b_0 - 1), \quad \beta = \frac{\alpha}{2}.$$

Since

$$\begin{aligned}\Phi_{\theta_0}^{-1}(\beta) - \Phi_{\theta_1}^{-1}(\beta) &= (1-\beta)a_0 + \beta b_0 - [(1-\beta)a_1 + \beta b_1] \\ &= (1-\beta)(a_0 - a_1) + \beta(b_0 - b_1) \\ &= 1 - \beta + \beta = 1 > 0,\end{aligned}$$

the linear uncertainty distribution family $\{\mathcal{L}(a, b) : a < b\}$ is nonembedded for θ_0 at level α.

From Definition 2, we also know that a regular uncertainty distribution family $\{\Phi_\theta : \theta \in \Theta\}$ is embedded for $\theta_0 \in \Theta$ at level α if

$$\Phi_{\theta_0}^{-1}(\beta) \le \Phi_{\theta}^{-1}(\beta) \quad \text{and} \quad \Phi_{\theta}^{-1}(1-\beta) \le \Phi_{\theta_0}^{-1}(1-\beta) \tag{2}$$

for any $\theta \in \Theta$ and any β with $0 < \beta \le \alpha/2$. It is obvious that (2) is equivalent to

$$[\Phi_{\theta}^{-1}(\beta), \Phi_{\theta}^{-1}(1-\beta)] \subseteq [\Phi_{\theta_0}^{-1}(\beta), \Phi_{\theta_0}^{-1}(1-\beta)],$$

which is the reason why $\{\Phi_\theta : \theta \in \Theta\}$ is named as embedded uncertainty distribution family. To illustrate the concept of embedded uncertainty distribution family, some examples are given as follows.

Example 3 The uncertainty distribution family

$$\left\{\mathcal{N}\left(0, \exp\left(-(\theta-1)^2\right)\right) : \theta \in \Re\right\}$$

is embedded for $\theta_0 = 1$ at any level α. Note that the inverse uncertainty distribution of

$$\mathcal{N}\left(0, \exp\left(-(\theta-1)^2\right)\right)$$

is

$$\Phi_{\theta}^{-1}(\beta) = 0 + \frac{\sqrt{3}}{\pi}\exp\left(-(\theta-1)^2\right)\ln\frac{\beta}{1-\beta} = \frac{\sqrt{3}}{\pi}\exp\left(-(\theta-1)^2\right)\ln\frac{\beta}{1-\beta}.$$

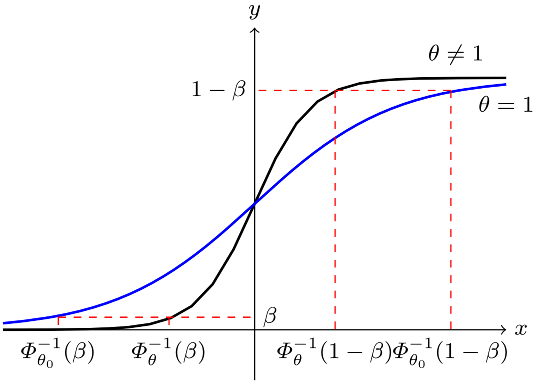

Fig. 1 The sketch map in Example 4

For any $\theta \in \Re$ and any β with $0 < \beta \le \alpha/2 < 0.5$, since

$$\begin{aligned}
&\Phi_{\theta_0}^{-1}(\beta) - \Phi_{\theta}^{-1}(\beta)\\
&= \frac{\sqrt{3}}{\pi}\exp\left(-(\theta_0-1)^2\right)\ln\frac{\beta}{1-\beta} - \frac{\sqrt{3}}{\pi}\exp\left(-(\theta-1)^2\right)\ln\frac{\beta}{1-\beta}\\
&= \frac{\sqrt{3}}{\pi}\left(\exp\left(-(\theta_0-1)^2\right) - \exp\left(-(\theta-1)^2\right)\right)\ln\frac{\beta}{1-\beta}\\
&= \frac{\sqrt{3}}{\pi}\left(\exp\left(-(1-1)^2\right) - \exp\left(-(\theta-1)^2\right)\right)\ln\frac{\beta}{1-\beta}\\
&= \frac{\sqrt{3}}{\pi}\left(1 - \exp\left(-(\theta-1)^2\right)\right)\ln\frac{\beta}{1-\beta} \le 0
\end{aligned}$$

and

$$\Phi_{\theta_0}^{-1}(1-\beta) - \Phi_{\theta}^{-1}(1-\beta) = \frac{\sqrt{3}}{\pi}\left(1 - \exp\left(-(\theta-1)^2\right)\right)\ln\frac{1-\beta}{\beta} \ge 0,$$

we have

$$\Phi_{\theta_0}^{-1}(\beta) \le \Phi_{\theta}^{-1}(\beta), \quad \Phi_{\theta}^{-1}(1-\beta) \le \Phi_{\theta_0}^{-1}(1-\beta),$$

which implies the uncertainty distribution family

$$\left\{\mathcal{N}\left(0, \exp\left(-(\theta-1)^2\right)\right) : \theta \in \Re\right\}$$

is embedded. A sketch map for ease of understanding is shown in Fig. 1.

Example 4 Whether an uncertainty distribution family is nonembedded is related to the value of θ_0 in Definition 2. For example, the uncertainty distribution family

$$\left\{\mathcal{N}\left(0, \exp\left(-(\theta-1)^2\right)\right) : \theta \in \Re\right\}$$

is embedded for $\theta_0 = 1$ at any level α, but nonembedded for any $\theta_0 \neq 1$ at any level α.

Example 5 Whether an uncertainty distribution family is nonembedded is also related to the level α in Definition 2. For example, for each $\theta \in \Re$, write

$$\Phi_\theta(x) = \begin{cases} 0, & \text{if } x \leq -0.2 \\ x + 0.2, & \text{if } -0.2 < x \leq 0 \\ 0.3\exp(\theta^2)x + 0.2, & \text{if } 0 < x \leq \exp(-\theta^2) \\ \dfrac{0.3}{2 - \exp(-\theta^2)}(x-2) + 0.8, & \text{if } \exp(-\theta^2) < x \leq 2 \\ x - 1.2, & \text{if } 2 < x \leq 2.2 \\ 1, & \text{if } x > 2.2. \end{cases}$$

Then, the uncertainty distribution family $\{\Phi_\theta : \theta \in \Re\}$ is nonembedded for any $\theta_0 \in \Re$ at any level α with $0.4 < \alpha < 1$, but embedded for any $\theta_0 \in \Re$ at any level α with $0 < \alpha \leq 0.4$.

Theorem 1 *Let ξ be a population with regular uncertainty distribution Φ_θ where θ is an unknown parameter with $\theta \in \Theta$. If the uncertainty distribution family $\{\Phi_\theta : \theta \in \Theta\}$ is nonembedded for a known parameter $\theta_0 \in \Theta$ at significance level α, then the test for the two-sided hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ at significance level α is*

$$W = \Bigg\{(z_1, z_2, \cdots, z_n) : \textit{ there are at least } \alpha \textit{ of indexes } i\textit{'s with } 1 \leq i \leq n \\ \textit{such that } z_i < \Phi_{\theta_0}^{-1}\left(\frac{\alpha}{2}\right) \textit{ or } z_i > \Phi_{\theta_0}^{-1}\left(1 - \frac{\alpha}{2}\right)\Bigg\}.$$

Proof In order to prove that W is a test for the two-sided hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ at level α, we need to verify that W satisfies the two conditions in Definition 1.

First, we will verify the condition (a) in Definition 1. For any $(z_1, z_2, \cdots, z_n) \in W$, it follows from the definition of W that there are at least α of indexes i's with $1 \leq i \leq n$ such that

$$z_i < \Phi_{\theta_0}^{-1}\left(\frac{\alpha}{2}\right) \quad \text{or} \quad z_i > \Phi_{\theta_0}^{-1}\left(1 - \frac{\alpha}{2}\right),$$

i.e.,

$$\mathcal{M}_{\theta_0}\{\xi > z_i\} > 1 - \frac{\alpha}{2} \quad \text{or} \quad \mathcal{M}_{\theta_0}\{\xi < z_i\} > 1 - \frac{\alpha}{2}.$$

Therefore W satisfies the condition (a).

Second, we will verify the condition (b). Since the uncertainty distribution family $\{\Phi_\theta : \theta \in \Theta\}$ is nonembedded for θ_0 at level α, we have

$$\Phi_{\theta_0}^{-1}(\beta) > \Phi_\theta^{-1}(\beta) \quad \text{or} \quad \Phi_\theta^{-1}(1-\beta) > \Phi_{\theta_0}^{-1}(1-\beta)$$

for some $\theta \in \Theta$ and some β with $0 < \beta \le \alpha/2$. Take

$$z_i = \begin{cases} \Phi_\theta^{-1}(\beta), & \text{if } \Phi_{\theta_0}^{-1}(\beta) > \Phi_\theta^{-1}(\beta) \text{ and } \Phi_\theta^{-1}(1-\beta) \le \Phi_{\theta_0}^{-1}(1-\beta) \\ \Phi_\theta^{-1}(1-\beta), & \text{if } \Phi_\theta^{-1}(1-\beta) > \Phi_{\theta_0}^{-1}(1-\beta), \end{cases}$$

$i = 1, 2, \cdots, n$. It is easy to verify that

$$\mathcal{M}_\theta\{\xi > z_i\} \vee \mathcal{M}_\theta\{\xi < z_i\} \le 1 - \beta$$

and

$$\mathcal{M}_{\theta_0}\{\xi > z_i\} \vee \mathcal{M}_{\theta_0}\{\xi < z_i\} > 1 - \beta, \tag{3}$$

$i = 1, 2, \cdots, n$. Thus,

$$\mathcal{M}_\theta\{\xi > z_i\} \vee \mathcal{M}_\theta\{\xi < z_i\} < \mathcal{M}_{\theta_0}\{\xi > z_j\} \vee \mathcal{M}_{\theta_0}\{\xi < z_j\},\ i, j = 1, 2, \cdots, n.$$

In addition, since $\beta \le \alpha/2$, it follows from (3) that

$$\mathcal{M}_{\theta_0}\{\xi > z_i\} \vee \mathcal{M}_{\theta_0}\{\xi < z_i\} > 1 - \beta \ge 1 - \frac{\alpha}{2},\ i = 1, 2, \cdots, n.$$

That is, $(z_1, z_2, \cdots, z_n) \in W$. Therefore W satisfies the condition (b). The theorem is proved. □

Remark 2 In order to make it easier to determine if the vector of observed data $(z_1, z_2, \cdots, z_n)$ falls into the test W defined in Theorem 1, we introduce a concept of singular point. For each i with $1 \le i \le n$, if

$$z_i < \Phi_{\theta_0}^{-1}\left(\frac{\alpha}{2}\right) \quad \text{or} \quad z_i > \Phi_{\theta_0}^{-1}\left(1 - \frac{\alpha}{2}\right),$$

then z_i is called a singular point. It follows from Theorem 1 that $(z_1, z_2, \cdots, z_n) \in W$ iff the number of singular points is at least αn, and $(z_1, z_2, \cdots, z_n) \notin W$ iff the number of singular points is less than αn.

Example 6 The condition of nonembedded uncertainty distribution family in Theorem 1 cannot be removed. For example, let ξ be a population with uncertainty distribution

$$\mathcal{N}\left(0, \exp\left(-(\theta - 1)^2\right)\right)$$

where θ is an unknown parameter. Write $\theta_0 = 1$. For a given significance level α, take the set

$$W = \left\{(z_1, z_2, \cdots, z_n) : \text{ there are at least } \alpha \text{ of indexes } i\text{'s with } 1 \le i \le n \right.$$
$$\left. \text{such that } z_i < \Phi_{\theta_0}^{-1}\left(\frac{\alpha}{2}\right) \text{ or } z_i > \Phi_{\theta_0}^{-1}\left(1 - \frac{\alpha}{2}\right)\right\}$$

where $\Phi_{\theta_0}^{-1}$ is the inverse uncertainty distribution of

$$\mathcal{N}\left(0, \exp\left(-(\theta_0 - 1)^2\right)\right).$$

It follows from the proof of Theorem 1 that the set W satisfies the condition (a) in Definition 1. However, we claim that the set W does not satisfy the condition (b) in Definition 1. To prove it, we employ the method of proof by contradiction. Suppose, on the contrary, that W satisfies the condition (b) in Definition 1. Then for some $\theta \ne \theta_0$ and some $(z_1, z_2, \cdots, z_n) \in W$, there are more than $1 - \alpha$ of indexes i's with $1 \le i \le n$ and at least α of indexes j's with $1 \le j \le n$ such that

$$\mathcal{M}_\theta\{\xi > z_i\} \vee \mathcal{M}_\theta\{\xi < z_i\} < \mathcal{M}_{\theta_0}\{\xi > z_j\} \vee \mathcal{M}_{\theta_0}\{\xi < z_j\},$$

i.e.,

$$\mathcal{M}_\theta\{\xi > z_i\} \vee \mathcal{M}_\theta\{\xi < z_i\} \le 1 - \beta < \mathcal{M}_{\theta_0}\{\xi > z_j\} \vee \mathcal{M}_{\theta_0}\{\xi < z_j\}$$

for some β with $0 < \beta \le \alpha/2$. Thus there exists an index k such that

$$\Phi_\theta^{-1}(\beta) \le z_k \le \Phi_\theta^{-1}(1 - \beta)$$

and

$$z_k < \Phi_{\theta_0}^{-1}(\beta) \quad \text{or} \quad z_k > \Phi_{\theta_0}^{-1}(1 - \beta).$$

Hence

$$\Phi_{\theta_0}^{-1}(\beta) > \Phi_\theta^{-1}(\beta) \quad \text{or} \quad \Phi_\theta^{-1}(1 - \beta) > \Phi_{\theta_0}^{-1}(1 - \beta),$$

which indicates that the uncertainty distribution family

$$\left\{\mathcal{N}\left(0, \exp\left(-(\theta - 1)^2\right)\right) : \theta \in \Re\right\}$$

is nonembedded for θ_0 at level α. This contradicts the conclusion shown in Example 3, i.e., the uncertainty distribution family

$$\left\{\mathcal{N}\left(0, \exp\left(-(\theta - 1)^2\right)\right) : \theta \in \Re\right\}$$

is embedded for θ_0 at level α. Thus W does not satisfy the condition (b) in Definition 1. Therefore the condition of nonembedded uncertainty distribution family cannot be removed.

Corollary 1 *Let ξ be a population that follows a normal uncertainty distribution with unknown expected value e and variance σ^2. Then the test for the two-sided hypotheses*

$$H_0 : e = e_0 \text{ and } \sigma = \sigma_0 \text{ versus } H_1 : e \neq e_0 \text{ or } \sigma \neq \sigma_0 \tag{4}$$

at significance level α is

$$\begin{aligned} W = \Big\{&(z_1, z_2, \cdots, z_n) : \text{ there are at least } \alpha \text{ of indexes } i\text{'s with } 1 \leq i \leq n \\ &\text{such that } z_i < \Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ or } z_i > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\Big\} \end{aligned} \tag{5}$$

where

$$\Phi^{-1}(\alpha) = e_0 + \frac{\sigma_0\sqrt{3}}{\pi} \ln \frac{\alpha}{1-\alpha}.$$

Proof Since Example 1 shows that the normal uncertainty distribution family

$$\{\mathcal{N}(e, \sigma) : e \in \Re, \sigma > 0\}$$

is nonembedded for (e_0, σ_0) at any significance level α, it follows from Theorem 1 that the test for hypotheses (4) is W defined in (5). □

Example 7 Let ξ be a population, and let $(z_1, z_2, \cdots, z_n)$ be a vector of observed data. In order to test whether ξ follows the normal uncertainty distribution $\mathcal{N}(e_0, \sigma_0)$, we may consider the two-sided hypotheses

$$H_0 : e = e_0 \text{ and } \sigma = \sigma_0 \text{ versus } H_1 : e \neq e_0 \text{ or } \sigma \neq \sigma_0. \tag{6}$$

Given a significance level α, it follows from Corollary 1 that the test for the hypotheses (6) at level α is

$$\begin{aligned} W = \Big\{&(z_1, z_2, \cdots, z_n) : \text{ there are at least } \alpha \text{ of indexes } i\text{'s with } 1 \leq i \leq n \\ &\text{such that } z_i < \Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ or } z_i > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\Big\} \end{aligned}$$

where

$$\Phi^{-1}(\alpha) = e_0 + \frac{\sigma_0\sqrt{3}}{\pi} \ln \frac{\alpha}{1-\alpha}.$$

If $(z_1, z_2, \cdots, z_n) \in W$, then we reject H_0. Otherwise, we accept H_0.

3 Uncertain regression analysis

In this section, we will apply the uncertain hypothesis test in uncertain regression analysis. Let $(x_1, x_2, \cdots, x_p)$ be a vector of explanatory variables, and let y be a response variable. Yao and Liu (2018) suggested that the functional relationship between $(x_1, x_2, \cdots, x_p)$ and y is expressed by an uncertain regression model

$$y = f(x_1, x_2, \cdots, x_p|\boldsymbol{\beta}) + \varepsilon$$

where $\boldsymbol{\beta}$ is a vector of parameters, and ε is an uncertain disturbance term (uncertain variable).

Suppose there is a set of observed data,

$$(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i),\ i = 1, 2, \cdots, n.$$

By employing least squares method (Yao and Liu (2018)), least absolute deviations method (Liu and Yang (2020)) or maximum likelihood method (Lio and Liu (2020)), we can obtain an estimation $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$. Then the fitted regression model is determined by

$$y = f(x_1, x_2, \cdots, x_p|\hat{\boldsymbol{\beta}}). \tag{7}$$

For each i $(i = 1, 2, \cdots, n)$, the i-th residual is

$$\varepsilon_i = y_i - f(x_{i1}, x_{i2}, \cdots, x_{ip}|\hat{\boldsymbol{\beta}}).$$

The residuals $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ can be regarded as the samples of the uncertain disturbance term ε. Thus, Lio and Liu (2018) suggested that the expected value of the uncertain disturbance term ε can be estimated as the average of residuals, i.e.,

$$\hat{e} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$$

and the variance can be estimated as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \hat{e})^2.$$

Therefore, we may assume the estimated disturbance term $\hat{\varepsilon}$ follows the normal uncertainty distribution $\mathcal{N}(\hat{e}, \hat{\sigma})$. Then the forecast uncertain variable of response variable y with respect to $(x_1, x_2, \cdots, x_p)$ is determined by

$$\hat{y} = f(x_1, x_2, \cdots, x_p|\hat{\boldsymbol{\beta}}) + \hat{\varepsilon},\ \hat{\varepsilon} \sim \mathcal{N}(\hat{e}, \hat{\sigma}).$$

In order to test whether the estimated disturbance term $\hat{\varepsilon}$ is appropriate, we consider the following hypotheses:

$$H_0 : e = \hat{e} \text{ and } \sigma = \hat{\sigma} \text{ versus } H_1 : e \neq \hat{e} \text{ or } \sigma \neq \hat{\sigma}. \tag{8}$$

Given a level of significance α (e.g. 0.05), it follows from Corollary 1 that the test for the hypotheses (8) is

$$\begin{aligned} W = \Big\{ &(z_1, z_2, \cdots, z_n) : \text{ there are at least } \alpha \text{ of indexes } i\text{'s with } 1 \leq i \leq n \\ &\text{such that } z_i < \Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ or } z_i > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \Big\} \end{aligned} \tag{9}$$

where

$$\Phi^{-1}(\alpha) = \hat{e} + \frac{\hat{\sigma}\sqrt{3}}{\pi} \ln \frac{\alpha}{1 - \alpha}.$$

For each i $(i = 1, 2, \cdots, n)$, if

$$\varepsilon_i < \Phi^{-1}\left(\frac{\alpha}{2}\right) \quad \text{or} \quad \varepsilon_i > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

then $(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i)$ is regarded as an outlier. If the number of outliers is at least αn, i.e.,

$$(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) \in W,$$

then either the estimated disturbance term $\mathcal{N}(\hat{e}, \hat{\sigma})$ or the fitted regression model (7) is inappropriate. Otherwise, both the estimated disturbance term $\mathcal{N}(\hat{e}, \hat{\sigma})$ and the fitted regression model (7) are appropriate.

4 Numerical Examples

This section will provide two examples to illustrate how to employ uncertain hypothesis test in uncertain regression analysis to test whether the estimated disturbance term and the fitted regression model are appropriate.

Example 8 Assume there is a set of observed data $(x_{i1}, x_{i2}, x_{i3}, y_i)$, $i = 1, 2, \cdots, 30$. See Table 1. In order to fit these observed data, we employ the linear uncertain regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where $\beta_0, \beta_1, \beta_2, \beta_3$ are some parameters, and ε is an uncertain disturbance term (uncertain variable).

Table 1 Observed data in Example 8

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_{i1}	4	6	9	4	5	6	7	5	9	5	6	7	4	9	10
x_{i2}	15	16	20	20	17	19	14	18	16	17	17	20	16	16	14
x_{i3}	21	21	24	26	28	20	23	26	25	18	29	30	21	29	27
y_i	45	50	61	52	54	48	52	57	56	48	55	61	53	56	59

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_{i1}	8	9	5	6	8	7	6	6	9	10	7	5	5	6	5
x_{i2}	19	15	20	15	16	20	20	16	18	17	18	14	15	18	17
x_{i3}	21	20	19	25	26	30	22	22	23	20	25	18	20	20	29
y_i	60	49	50	47	54	59	56	46	61	50	58	38	44	50	54

Using the observed data in Table 1 and solving the minimization problem

$$\min_{\beta_0,\beta_1,\beta_2,\beta_3} \sum_{i=1}^{30} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2,$$

we obtain the fitted linear regression model

$$y = 4.3965 + 1.3644 x_1 + 1.3130 x_2 + 0.7166 x_3. \tag{10}$$

From

$$\varepsilon_i = y_i - 4.3965 - 1.3644 x_{i1} - 1.3130 x_{i2} - 0.7166 x_{i3},\ i = 1, 2, \cdots, 30,$$

we obtain 30 residuals $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{30}$. Thus the expected value of estimated disturbance term $\hat{\varepsilon}$ is

$$\hat{e} = \frac{1}{30} \sum_{i=1}^{30} \varepsilon_i = 0.0000,$$

and the variance is

$$\hat{\sigma}^2 = \frac{1}{30} \sum_{i=1}^{30} (\varepsilon_i - \hat{e})^2 = 2.8529^2.$$

Therefore, we may assume the estimated disturbance term $\hat{\varepsilon}$ follows the normal uncertainty distribution $\mathcal{N}(0.0000, 2.8529)$. Then the forecast uncertain variable of response variable y with respect to (x_1, x_2, x_3) is determined by

$$\hat{y} = 4.3965 + 1.3644 x_1 + 1.3130 x_2 + 0.7166 x_3 + \hat{\varepsilon},\ \hat{\varepsilon} \sim \mathcal{N}(0.0000, 2.8529).$$

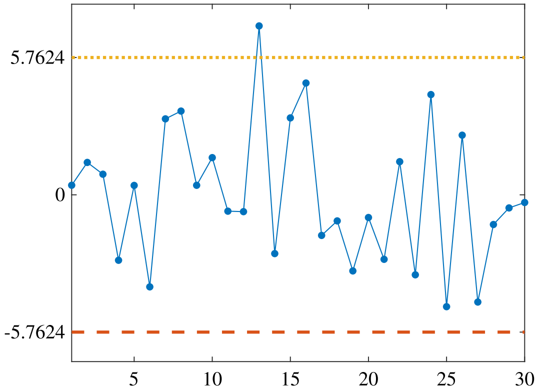

Fig. 2 Residual plot in Example 8

To test whether $\mathcal{N}(0.0000, 2.8529)$ is appropriate, we consider the following hypotheses:

$$H_0 : e = 0.0000 \text{ and } \sigma = 2.8529 \text{ versus } H_1 : e \neq 0.0000 \text{ or } \sigma \neq 2.8529. \quad (11)$$

Given a significance level $\alpha = 0.05$, we obtain

$$\Phi^{-1}\left(\frac{\alpha}{2}\right) = -5.7624, \quad \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = 5.7624$$

where Φ^{-1} is the inverse uncertainty distribution of $\mathcal{N}(0.0000, 2.8529)$, i.e.,

$$\Phi^{-1}(\alpha) = 0.0000 + \frac{2.8529\sqrt{3}}{\pi} \ln \frac{\alpha}{1-\alpha}.$$

Since $\alpha \times 30 = 1.5$, it follows from (9) that the test for the hypotheses (11) is

$$W = \{(z_1, z_2, \cdots, z_{30}) : \text{ there are at least 2 of indexes } i\text{'s with } 1 \leq i \leq 30 \\ \text{such that } z_i < -5.7624 \text{ or } z_i > 5.7624\}.$$

As shown in Fig. 2, we can see that only

$$\varepsilon_{24} \notin [-5.7624, 5.7624].$$

Thus $(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{30}) \notin W$. Therefore we think both the estimated disturbance term $\mathcal{N}(0.0000, 2.8529)$ and the fitted linear regression model (10) are appropriate.

Example 9 Assume there is a set of observed data $(x_{i1}, x_{i2}, x_{i3}, y_i), i = 1, 2, \cdots, 30$. See Table 2. In order to fit these observed data, we employ the linear uncertain regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Table 2 Observed data in Example 9

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_{i1}	4	5	9	7	4	6	7	4	6	8	8	7	9	9	10
x_{i2}	17	16	15	18	16	16	16	15	19	15	15	16	20	17	16
x_{i3}	21	21	22	28	23	22	30	28	22	18	20	23	30	23	21
y_i	44	48	56	56	49	48	58	50	59	47	48	52	62	53	52

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_{i1}	4	10	8	5	10	10	4	9	5	7	6	10	4	6	6
x_{i2}	17	14	14	20	17	18	20	18	19	19	18	16	20	20	14
x_{i3}	27	20	28	27	26	26	30	19	23	26	19	23	28	19	23
y_i	60	47	56	55	65	59	58	54	51	57	50	52	60	50	46

where $\beta_0, \beta_1, \beta_2, \beta_3$ are some parameters, and ε is an uncertain disturbance term (uncertain variable).

Using the observed data in Table 2 and solving the minimization problem

$$\min_{\beta_0,\beta_1,\beta_2,\beta_3} \sum_{i=1}^{30}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2,$$

we obtain the fitted linear regression model

$$y = 4.5285 + 1.0549x_1 + 1.1399x_2 + 0.9292x_3. \tag{12}$$

From

$$\varepsilon_i = y_i - 4.5285 - 1.0549x_{i1} - 1.1399x_{i2} - 0.9292x_{i3},$$

we obtain 30 residuals $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{30}$. Thus the expected value of estimated disturbance term $\hat{\varepsilon}$ is

$$\hat{e} = \frac{1}{30}\sum_{i=1}^{30} \varepsilon_i = 0.0000,$$

and the variance is

$$\hat{\sigma}^2 = \frac{1}{30}\sum_{i=1}^{30}(\varepsilon_i - \hat{e})^2 = 2.7449^2.$$

Therefore, we may assume the estimated disturbance term $\hat{\varepsilon}$ follows the normal uncertainty distribution $\mathcal{N}(0.0000, 2.7449)$. Then the forecast uncertain variable of response

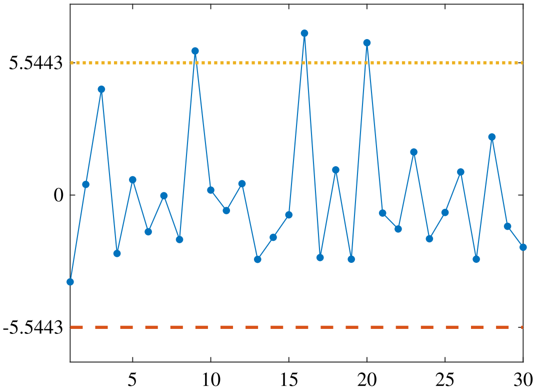

Fig. 3 Residual plot in Example 9

variable y with respect to (x_1, x_2, x_3) is determined by

$$\hat{y} = 4.5285 + 1.0549x_1 + 1.1399x_2 + 0.9292x_3 + \hat{\varepsilon},\ \hat{\varepsilon} \sim \mathcal{N}(0.0000, 2.7449).$$

To test whether $\mathcal{N}(0.0000, 2.7449)$ is appropriate, we consider the following hypotheses:

$$H_0 : e = 0.0000 \text{ and } \sigma = 2.8529 \text{ versus } H_1 : e \neq 0.0000 \text{ or } \sigma \neq 2.8529. \quad (13)$$

Given a significance level $\alpha = 0.05$, we obtain

$$\Phi^{-1}\left(\frac{\alpha}{2}\right) = -5.5443, \quad \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = 5.5443,$$

where Φ^{-1} is the inverse uncertainty distribution of $\mathcal{N}(0.0000, 2.7449)$, i.e.,

$$\Phi^{-1}(\alpha) = 0.0000 + \frac{2.7449\sqrt{3}}{\pi} \ln \frac{\alpha}{1-\alpha}.$$

Since $\alpha \times 30 = 1.5$, it follows from (9) that the test for the hypotheses (13) is

$$W = \{(z_1, z_2, \cdots, z_{30}) : \text{ there are at least 2 of indexes } i\text{'s with } 1 \leq i \leq 30$$
$$\text{such that } z_i < -5.5443 \text{ or } z_i > 5.5443\}.$$

As shown in Fig. 3, we can see that

$$\varepsilon_{10} > 5.5443, \ \varepsilon_{12} > 5.5443, \ \varepsilon_{15} > 5.5443.$$

Thus $(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{30}) \in W$. Therefore we think either the estimated disturbance term $\mathcal{N}(0.0000, 2.7449)$ or the fitted linear regression model (12) is inappropriate.

5 Conclusion

This paper first introduced a mathematical tool of uncertain hypothesis test to decide whether some hypotheses are correct or not, based on observed data. With the help of the concept of nonembedded uncertainty distribution family, the test for two-sided hypotheses was constructed. Then uncertain hypothesis test was employed in uncertain regression analysis to test whether the estimated disturbance term and the fitted regression model are appropriate. Finally, this paper gave some numerical examples to illustrate the test process.

In the future, the uncertain hypothesis test will be applied in other development fields of uncertain statistics like estimating uncertainty distribution, uncertain time series analysis and parameter estimation in uncertain differential equation.

Acknowledgements This work was supported by the National Natural Science Foundation of China Grant No.61873329 and Grant No.12026225.

References

Chen, X., & Ralescu, D. (2012). B-Spline method of uncertain statistics with application to estimating travel distance. *Journal of Uncertain Systems*, *6*(4), 256–262.

Ding, J., & Zhang, Z. (2021). Statistical inference on uncertain nonparametric regression model. *Fuzzy Optimization and Decision Making*. https://doi.org/10.1007/s10700-021-09353-0.

Lio, W., & Liu, B. (2018). Residual and confidence interval for uncertain regression model with imprecise observations. *Journal of Intelligent & Fuzzy Systems*, *35*(2), 2573–2583.

Lio, W., & Liu, B. (2020). Uncertain maximum likelihood estimation with application to uncertain regression analysis. *Soft Computing*, *24*, 9351–9360.

Lio, W., & Liu, B. (2021). Initial value estimation of uncertain differential equations and zero-day of COVID-19 spread in China. *Fuzzy Optimization and Decision Making*, *20*(2), 177–188.

Liu, B. (2007). *Uncertainty Theory* (2nd ed.). Berlin: Springer.

Liu, B. (2009). Some research problems in uncertainty theory. *Journal of Uncertain Systems*, *3*(1), 3–10.

Liu, B. (2010). *Uncertainty Theory: A Branch of Mathematics for Modeling Human Uncertainty*. Berlin: Springer.

Liu, S. (2019). Leave-*p*-out cross-validation test for uncertain Verhulst-Pearl model with imprecise observations. *IEEE Access*, *7*, 131705–131709.

Liu, Y., & Liu, B. (2020). Estimating unknown parameters in uncertain differential equation by maximum likelihood estimation. Technical Report.

Liu, Z. (2021). Generalized moment estimation for uncertain differential equations. *Applied Mathematics and Computation*, *392*, 125724.

Liu, Z., & Jia, L. (2020). Cross-validation for the uncertain Chapman-Richards growth model with imprecise observations. *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems*, *5*(28), 769–783.

Liu, Z., & Yang, X. (2020). Variable selection in uncertain regression analysis with imprecise observations. Technical Report.

Liu, Z., & Yang, Y. (2020). Least absolute deviations estimation for uncertain regression with imprecise observations. *Fuzzy Optimization and Decision Making*, *19*(1), 33–52.

Sheng, Y. H., Yao, K., & Chen, X. (2020). Least squares estimation in uncertain differential equations. *IEEE Transactions on Fuzzy Systems*, *28*(10), 2651–2655.

Song, Y., & Fu, Z. (2018). Uncertain multivariable regression model. *Soft Computing*, *22*(17), 5861–5866.

Wang, X., Gao, Z., & Guo, H. (2012). Delphi method for estimating uncertainty distributions. *Information: An International Interdisciplinary Journal*, *15*(2), 449–460.

Wang, X., & Peng, Z. (2014). Method of moments for estimating uncertainty distributions. *Journal of Uncertainty Analysis and Applications*, *2*, 5.

Yang, X., & Liu, B. (2019). Uncertain time series analysis with imprecise observations. *Fuzzy Optimization and Decision Making*, *18*(3), 263–278.

Yang, X., Liu, Y., & Park, G. (2020). Parameter estimation of uncertain differential equation with application to financial market. *Chaos, Solitons and Fractals*, *139*, 110026.

Yang, X., & Ni, Y. (2020). Least-squares estimation for uncertain moving average model. *Communications in Statistics-Theory and Methods*. https://doi.org/10.1080/03610926.2020.1713373.

Yao, K., & Liu, B. (2018). Uncertain regression analysis: An approach for imprecise observations. *Soft Computing*, *22*(17), 5579–5582.

Yao, K., & Liu, B. (2020). Parameter estimation in uncertain differential equations. *Fuzzy Optimization and Decision Making*, *19*(1), 1–12.

Ye, T., & Liu, Y. (2020). Multivariate uncertain regression model with imprecise observations. *Journal of Ambient Intelligence and Humanized Computing*, *11*, 4941–4950.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.